Técnicas de clustering aplicadas en un conjunto metabolitos perteneciente a pacientes con Leishmaniasis cutánea para predecir la efectividad del tratamiento Glucantime a través de modelos de aprendizaje automático clásico

Juan Pablo Luna Mejía & Naim Samuel Sadeghian Perskie

Resumen

Los medicamentos usados para el tratamiento de la leishmaniasis pueden ser tóxicos y detrimentales para la salud. Peor aún, estos tratamientos no prometen curar al paciente en todos los casos. Para evitar recetar dichos tratamientos a pacientes a los que no les va a funcionar el manejo, se han hecho varios estudios para tratar de predecir, por medio de muestras de metabolitos en la sangre, en qué pacientes el tratamiento será efectivo. En este proyecto se hizo una continuación de estos estudios, basados en los mismos datos usados. Estos datos observaron 535 atributos/metabolitos para solo 36 pacientes. El grueso de este proyecto se centraba en reducir la dimensionalidad del conjunto de datos (2 a 5 metabolitos) y poder llegar a resultados cercanos o mucho mejores a los ya existentes. Se entrenaron 4 diferentes modelos de agrupamiento para encontrar posibles grupos y de cada uno escoger un representante. Para cada modelo se buscaron los parámetros de los cuales llegaban a clúster con un mejor grado de separación. En la fase de escoger los representantes de cada clúster se usaron diferentes métricas como: cercanía al centro del clúster o probabilidad de ser miembro del clúster, para así decidir cuáles podrían ser los mejores representantes. Después de la fase mencionada anteriormente se dio paso a la fase de predicción, donde se observó qué tan buena era la predicción con este pequeño conjunto de atributos. Finalmente se llegó a un modelo con 3 metabolitos y un puntaje f1 de 0.82, el cual fue muy prometedor para una forma de reducción de la dimensionalidad tan particular.

Introducción

Este proyecto parte de un conjunto de datos obtenidos gracias a una investigación previa sobre el funcionamiento del tratamiento para la leishmaniasis [1]. En este estudio se trató de predecir el dicho funcionamiento mirando un perfil de los metabolitos en la sangre de cada paciente y registrando si el paciente se vio beneficiado o no. La misión de este estudio será hacer predicciones sobre este conjunto, pero esta vez con aprendizaje automático. El conjunto cuenta con un problema en cuanto a su dimensionalidad: observa 536 atributos, pero solo tiene 35 registros. Estas características son poco deseables para el aprendizaje automático, el cual se beneficia de conjuntos de muchos registros y pocos atributos. Por lo tanto, el objetivo del proyecto será seleccionar un conjunto de datos utilizando algoritmos de agrupamiento, para poder hacer predicción con modelos de aprendizaje automático y a su vez encontrar mejores y más eficientes resultados. El grueso de este proyecto se llevará a cabo en dos etapas principales. La primera, será la reducción de la dimensionalidad; donde se seleccionarán entrenan diferentes modelos de agrupamiento para dividir los metabolitos en subconjuntos con características similares. Una vez se encuentren estos grupos que forman los

metabolitos, se tomará un miembro de cada grupo como representante del resto; para así, reducir el ancho de los atributos de 536 a unos 2 a 5. La segunda etapa, ya será el proceso de predicción donde se probará hacer predicción usando la tabla con metabolitos escogidos en la etapa anterior como atributos. Finalmente, se recopilarán los resultados para ver si tienen algún uso en el campo de acción y se los resultados son mejores que los del estudio anterior.

Fundamentación teórica

Para el desarrollo del proyecto se utilizaron 4 algoritmos de agrupamiento y 3 algoritmos de predicción. Aparte de esto, se utilizaron varias métricas de medición, principalmente la Sihouette Score y el método de estandarización de datos de Pareto.

La estandarización de Pareto es simple, consiste en reemplazar los valores de cada columna por el valor inicial dividido entre la raíz cuadrada de la desviación estándar de la columna [2].

El Silhouette Score es una métrica de la calidad de un algoritmo de agrupamiento en la que se evalúa qué tan bien los objetos están agrupados. En general, el Silhouette Score se utiliza para determinar el número óptimo de clústeres en un conjunto de datos y que tan bien están conformados estos.

Los 4 algoritmos de agrupamiento que se utilizaron fueron K-means, DBSCAN, mixturas gaussianas y algoritmo jerárquico aglomerativo.

K-means es una técnica de aprendizaje no supervisado que se utiliza para agrupar datos en diferentes grupos o clústeres. El objetivo del algoritmo es encontrar k centroides, donde k es el número de grupos o clústeres que se desea crear, y luego asignar cada punto de datos al centroide más cercano. DBSCAN es una técnica de agrupamiento que, a diferencia del algoritmo K-means, es sensible a la elección del número de clústeres deseados. DBSCAN determina automáticamente el número de clústeres y también puede identificar puntos que no pertenecen a ningún clúster. El algoritmo de mixturas gaussianas por otra parte es una técnica de aprendizaje no supervisado que se utiliza para modelar la distribución de probabilidad de un conjunto de datos y para identificar los diferentes grupos o clústeres que conforman ese conjunto de datos. El algoritmo jerárquico aglomerativo es un método de agrupamiento que se utiliza para dividir un conjunto de datos en grupos o clústeres en función de su similitud. A diferencia de los algoritmos de agrupamiento de partición, que dividen el conjunto de datos en un número fijo de clústeres, el algoritmo jerárquico aglomerativo construye una jerarquía de clústeres anidados.

Los modelos de predicción clásicos que se implementaron fueron 3. Estos 3 fueron regresión lineal, K-vecinos y Random Forest.

Regresión lineal es utilizado para predecir el valor de una variable numérica continúa basándose en una o varias variables de entrada. El objetivo de la regresión lineal es encontrar la mejor línea recta que se ajuste a los datos y que permita predecir valores nuevos con la menor cantidad de error posible. K-vecinos clasifica o predice el valor de una variable numérica basándose en las características de sus vecinos más cercanos. El valor de K representa el número de vecinos más cercanos que se utilizan para hacer la predicción, y se eligen los que están más cerca del punto a clasificar en un espacio de características definido por los datos. Random Forest se basa en la creación de múltiples árboles de decisión y la combinación de sus predicciones para obtener una predicción final más precisa. Cada árbol se construye utilizando una muestra aleatoria de los datos y una

selección aleatoria de características, lo que ayuda a evitar el sobreajuste y a mejorar la generalización del modelo.

Resultados

Las áreas principales que se tocarán para este trabajo de investigación son tres: Primero está el preprocesamiento de datos donde se busca optimizar y limpiar esta data para posteriormente utilizarlos en la selección de metabolitos. La segunda en la selección de metabolitos donde se utilizan distintas técnicas de agrupamiento para agrupar los 535 metabolitos y seguidamente seleccionar los representantes de estos grupos. Por último, se observa la predicción de resultados.

Para el agrupamiento de metabolitos lo que se busca es elegir de los 535 que hay en la base de datos a solamente unos 5 o 6 metabolitos que estén fuertemente relacionados con la predicción de la recuperación del tratamiento contra la leishmaniasis. Para esto se van a utilizar algoritmos de agrupamiento, lo cual será el énfasis principal, para poner todos los 535 metabolitos en grupos. Luego de formar estos grupos se tomará el representante de cada uno de ellos y se evaluará que tan buen predictor es. Esto se hará con los algoritmos de predicción clásica.

Lo primero que se realizó fue preprocesar los datos para disminuir la dimensionalidad y la disparidad de los datos. Al contar con tantos datos y tanta disparidad esto era una parte clave para tener mejores resultados. Para reducir la dimensionalidad de los datos se eliminó un dato de cada par donde la correlación de este par de datos era del 87.5% o superior.

Para reducir la disparidad de los datos se utilizó la estandarización de Pareto. Este método ha probado ser útil basado en un trabajo anterior sobre esta misma base de datos [2], por lo que se decidió implementar este mismo método.

La segunda fase fue la fase de la selección de metabolitos que se haría con los metabolitos representantes de clúster formado por algoritmos de agrupamiento. Para la selección de dichos algoritmos el criterio que se eligió fue variedad. Seleccionar diferentes tipos de algoritmos culla naturaleza en como agrupan los datos es diferente de los demás algoritmos seleccionados y así tener resultados variados. De esto se seleccionaron 4 algoritmos K-means, DBSCAN, Mixturas gaussianas y algoritmo jerárquico aglomerativo. Para cada algoritmo lo que se hizo fue una grilla donde se encontraron los parámetros óptimos y consiguiente a ello cada uno de estos fue evaluado con diferentes scores, principalmente el Silhouette Score.

Para cada algoritmo se determinó un numero de clústers bajo, entre 3 y 6, donde normalmente había un cluster que tenía la gran mayoría de los metabolitos agrupados.

Posterior a esto se extrajeron los representantes de cada uno de los clústeres de cada algoritmo. Por ejemplo: se tomaba el algoritmo de K-means cuando generó 3 agrupaciones, luego de estas 3 agrupaciones se tomaba el centroide de cada clúster y se tomaba el metabolito más cercano al centroide y este sería el representante. Se repitió este proceso con todos los algoritmos con excepción de DBSCAN. Esto se decidió ya que DBSCAN en todos los casos, por su naturaleza y la forma por cómo estaban construidos nuestros datos, generaba un solo clúster enorme con mucho ruido.

Una vez se seleccionaron los metabolitos más representativos se pasó a la etapa de comprobar que tanta información brindan en cuanto a la predicción del funcionamiento del tratamiento. Se seleccionaron 3 modelos

de clasificación para observar los resultados: K-Nearest Neighbors, Logistic Regression y Random Forest. La implementación de estos algoritmos fue tomada de la librería de Sklearn [3]. Para cada uno de los modelos se hizo una búsqueda de grilla para encontrar los mejores parámetros. Además, se hizo un proceso de validación cruzada de K-Folds con 7 splits (grupos de 5) y 3 repeticiones. Al final de la búsqueda, para cada modelo se registró cuál fue la iteración con el mejor resultado comparando así el puntaje f1. Al final se obtuvieron una gran cantidad de resultados y algunos de estos bastante prometedores. En la siguiente tabla (tabla 1) se muestran los mejores resultados por representante de cada algoritmo de agrupamiento.

K-Medias		
Metodo	F1	Clusters
K-vecinos	0.6761905	3
Random Forest	0.8266667	3
Logistic Reg	0.7904762	5
Mixturas Gaussianas		
K-vecinos	0.6095238	2
Random Forest	0.8166667	2
Logistic Reg	0.6285714	2
Jerarquico Aglomerativo		
K-vecinos	0.6952381	3
Random Forest	0.7357143	3
Logistic Reg	0.782381	3

Tabla 1: Mejor resultado por algoritmos de agrupamiento

Discusión

Teniendo en cuenta de que las métricas resultan prometedoras y de que a su vez presentan puntajes que no son muy altos, se da a creer que sí puede haber una correlación entre la cantidad de un metabolito y el funcionamiento del tratamiento. Es difícil decir que los resultados de la predicción dan espacio a extraer conclusiones certeras o con suficiente peso. No sólo porque no es un puntaje muy alto, sino porque el conjunto de entrenamiento es muy poco; con solo unos 35 datos de la población. A pesar de esto, lo que sí ofrece este análisis es la facilidad de poder seguir llevando a cabo un entrenamiento del modelo con muy pocos parámetros. A comparación de los resultados de estudios anteriores, no destacan mucho los números que se lograron. El estudio también presenta otras etapas como la extracción de metabolitos altamente correlacionados, los cuales podrían proveer información interesante e importante.

Después de observar los resultados de DBSCAN, donde casi todos los metabolitos quedaron en un gran conjunto y el resto fueron tomados como ruido, se podría resultar de interés hacer un estudio teniendo en cuenta sólo este conjunto grande de metabolitos. Podría darse el caso de que existan subconjuntos que se pierden en los otros modelos, ya que los metabolitos del conjunto grande están tan cercanos entre sí. A futuro, también se podría hacer énfasis en la etapa de predicción utilizando modelos que no se tuvieron en cuenta en este estudio, pero sí en el de Zuluaga, tales como: Naive Bayes, Decision Trees o Support Vector Classification. Además, valdría la pena utilizar otros métodos de selección de atributos, los cuales no estén basados en el mejor representante de cada grupo, ya que hay una posibilidad de que los métodos usados para la selección de mejores representantes para cada uno de ellos no aseguren desempeño más eficiente en la fase de predicción. Es posible que los mejores metabolitos para la predicción no sean aquellos que representan un clúster, sino los que presentan características más únicas y se encuentran más alejados del resto.

Referencias

[1] Vargas DA, Prieto MD, Martínez-Valencia AJ, Cossio A, Burgess KEV, Burchmore RJS and Gómez MA (2019) Pharmacometabolomics of Meglumine Antimoniate in Patients With Cutaneous Leishmaniasis. Front. Pharmacol. 10:657. doi: 10.3389/fphar.2019.00657

[2] S. Zuluaga, “Aprendizaje de Máquina Aplicado a la Predicción del Éxito del Tratamiento de la Leishmaniasis,” Pontificia Universidad Javeriana Cali, 2022.

[3] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[4] Ministerio de Saludo, "GUÍA PARA LA ATENCIÓN CLÍNICA INTEGRAL DEL PACIENTE CON LEISHMANIASIS", Convenio de Cooperación Técnica con el Ministerio de la Protección Social Nro. 256 de 2023 y Nro. 237 de 2010

[5] G. Stegmayer, D. H. Milone, L. Kamenetzky, M. G. Lopez and F. Carrari, .A Biologically Inspired Validity Measure for Comparison of Clustering Methods over Metabolic Data Sets,ïn IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 3, pp. 706-716, May-June 2012, doi: 10.1109/TCBB.2012.10

[6] Bombardier, C.H., Divine, G.W., Jordan, J.S. et al. Minnesota Multiphasic Personality Inventory (MMPI) cluster groups among chronically ill patients: Relationship to illness adjustment and treatment outcome. J Behav Med 16, 467–484 (1993). https://doi.org/10.1007/BF00844817

[7] S. Hunta, N. Aunsri and T. Yooyativong, "Drug-Drug Interactions prediction from enzyme action crossing through machine learning approaches,"2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2015, pp. 1-4, doi: 10.1109/ECTICon.2015.7207126

[8] Ester, M., Kriegel, H. P., Sander, J., Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, No. 34, pp. 226-231).

[9] K. Arvai, "Knee LocatorÇopyright Revision 13b9c17d. 2020.