Optimización de clasificadores de arritmias mediante algoritmos genéticos

Universidad tecnológica nacional, Facultad Regional Rosario

Abstract
El propósito de este trabajo es demostrar cómo pueden optimizarse los clasificadores de arritmias utilizados (Gaussian Naive bayes y Knn) con el uso de algoritmos genéticos, a fin de poder clasificar con mayor precisión los tipos de arritmias presentes en un ECG.
Se han utilizado los ECG presentes en la base de datos de arritmias del MIT[1] y librerías para extraer y procesar datos de los mismos como wfdb y neurokit2 en Python.

Palabras claves:
ECG, Clasificador, Naive Bayes, KNN, Algoritmo Genético, Arritmia, matriz de confusión, mit bih arrhythmia database, enfermedades cardiovasculares.

Introducción:

Las arritmias son un trastorno de la frecuencia cardíaca (pulso) o del ritmo cardíaco. El corazón puede latir demasiado rápido, lo cual es denominado taquicardia, demasiado lento, denominado bradicardia o de manera irregular.

Una arritmia puede no causar daño, ser una señal de otros problemas cardíacos o un peligro inmediato para su salud.

El corazón no funciona de manera completamente predecible porque está controlado por un sistema de marcapasos natural (principalmente el nodo sinoauricular) y por señales del sistema nervioso autónomo, que incluyen tanto estímulos que lo aceleran (como el estrés o el ejercicio) como aquellos que lo ralentizan (como el descanso o la relajación). Además, las señales químicas, como las hormonas liberadas durante situaciones de estrés, también influyen en el ritmo cardíaco.

La actividad cardiaca es un sistema caótico y complejo, se comporta de manera dinámica presentando en algunos casos aspectos aleatorios y es sensible a las condiciones iniciales del sistema. Debido a la impredecibilidad mencionada, un ECG es tardado de clasificar manualmente, pero fundamental para reconocer arritmias de manera temprana, lo cual puede salvar millones de vidas en el mundo. El uso de clasificadores como los que veremos a continuación podrían llegar a facilitar el flujo de trabajo en el campo de estudio de las afecciones cardiacas.
El uso de algoritmos genéticos nos permiten modificar los clasificadores de tal manera que su eficacia sea mayor.

Desarrollo

Se ha optado por utilizar los clasificadores Naive Bayes y KNN. Primeramente estos clasificadores son entrenados con una parte de los registros de los electrocardiogramas

los cuales cuentan con una característica y una etiqueta asociada, para luego ser optimizados mediante algoritmos genéticos y comparar su rendimiento y eficacia.

Clasificador Naive Bayes:
Considerado un clasificador probabilístico, ya que se basa en el teorema de Bayes.

$$P(Y|X) = \frac{P(X \text{ and } Y)}{P(X)}$$

A diferencia de los clasificadores discriminatorios, como la regresión logística, no aprende qué características son más importantes para diferenciar entre clases.
Optamos por usar una variante del clasificador Naïve Bayes, llamado **Gaussian Naïve Bayes**, que se utiliza con distribuciones gaussianas, es decir distribuciones normales y variables continuas. Este modelo se ajusta encontrando la media y la desviación estándar de cada clase.

Clasificador KNN:
Es un clasificador de aprendizaje supervisado no paramétrico, que utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual.

A diferencia de otros algoritmos de aprendizaje supervisado, KNN no tiene una fase de entrenamiento como tal. En lugar de construir un modelo explícito, KNN simplemente almacena los datos de entrenamiento. Estos datos consisten en un conjunto de instancias, donde cada instancia tiene características (variables) y una etiqueta de clase (en el caso de clasificación) o un valor continuo (en el caso de regresión).

El valor k en el algoritmo KNN define cuántos vecinos se verificarán para determinar la clasificación de un punto de consulta específico.

Entrenamiento de los clasificadores:
Al leer cada registro seleccionado de la base de datos con el uso de la librería wfdb, se extraen sus características principales, tales como sus intervalos rr, la raíz cuadrada de la media de las diferencias cuadradas entre intervalos consecutivos de RR en la señal de ECG, la media aritmética y la desviación estándar.

Este conjunto de características forman un conjunto de datos representativo de cada ECG analizado. Por otro lado, extraemos las anotaciones correspondientes de esas características.

Terminado de leer todos los registros y extraídas las características y anotaciones tendremos dos conjuntos **x** e **y** respectivamente, que serán balanceados con el uso de SMOTE para no tener valores nulos.

Los conjuntos **x** e **y** se utilizaran para crear dos pares de conjuntos, uno que representara los datos de entrenamiento (80% de los elementos) y el otro los datos de prueba (20% restante de los elementos).

Ambos clasificadores se entrenan y probarán una vez con estos dos conjuntos, y luego se volverá a realizar el mismo entrenamiento y prueba una vez sean optimizados con algoritmos genéticos **(los datos de rendimiento de estas pruebas se**

encuentran en las tablas al final del documento).

Algoritmo Genético:
Los algoritmos genéticos son métodos sistemáticos para la resolución de problemas de búsqueda y de optimización de parámetros que aplican los mismos métodos de la evolución biológica: selección basada en población, mutación y reproducción sexual.

Este tipo de algoritmo requiere que cada individuo de la población se codifique en un cromosoma el cual tiene varios genes, que corresponden a sendos parámetros del problema.

Se define una función objetivo la cual debe encontrarse un mínimo o un máximo dependiendo el problema planteado y una función fitness que puede coincidir con la función objetivo, la cual medirá el potencial como padres de cada individuo.

Cada uno de los individuos de la población es evaluado en la **función fitness** determinando así una puntuación. La población es sometida a una selección en base a su potencial para determinar quiénes serán los padres de la próxima generación.

Función fitness de clasificador GaussianNB:

```
def fitness(individual):
    var_smoothing = 10**(-individual[0])
    clasificador = GaussianNB(var_smoothing=var_smoothing)
    clasificador.fit(X_train, y_train)
    y_pred = clasificador.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    return accuracy,
```

Función fitness de clasificador KNN:

```
def fitness(individual):
    k = int(individual[0])
    k = max(1, min(k, 20))  # Asegurar que k esté en el rango [1, 20]
    clasificador = KNeighborsClassifier(n_neighbors=k)
    clasificador.fit(X_train, y_train)
    y_pred = clasificador.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    return (accuracy,)
```

El crossover es el primordial operador genético y consiste en la cruza de duplas de cromosomas seleccionados, derivando en dos nuevos individuos.

La mutación consta de que el valor de uno de los genes del individuo varía de forma aleatoria, para este proyecto elegimos una **mutación polinomial acotada**. Dentro de esta clase de mutación el gen es alterado siguiendo una distribución polinómica, derivando en un individuo mutado que se encuentre dentro de una rango acotado

En la **selección** un operador genético extrae, aleatoriamente, pares de elementos de la población

El elitismo consiste en que los individuos más aptos (mejor puntuación fitness) pasen a la siguiente población. Esto garantiza la convergencia global del algoritmo genético

Aplicación en la optimización:
El algoritmo cuenta con una población inicial de 50 individuos, 40 corridas, 50% de probabilidad de crossover y un 20% de probabilidad de mutación.

Escogimos como método de selección el de torneo con elitismo. El torneo consta de la elección aleatoria de un número de individuos de la población, y el que cuenta con mayor puntuación se reproduce sustituyendo al de menor puntaje.

La función fitness en nuestro caso resulta de evaluar a cada individuo en los clasificadores.

Resultados para clasificador Gaussian:
Previo a la optimización con el algoritmo genético obtuvimos un acierto del 67,20% (Tabla 1)
Luego de optimizar el clasificador se pudo obtener un acierto del 67,45% (Tabla 2)

Resultados para clasificador KNN
Previo a la optimización con el algoritmo genético obtuvimos un acierto del 84,62% (Tabla 3).
Luego de optimizar el clasificador obtuvimos un acierto del 85,05% (Tabla 4)

Trabajos Relacionados
Este proyecto está basado en **"Algoritmos genéticos aplicados a la optimización de características en la clasificación de arritmias cardiacas utilizando los clasificadores KNN y naive Bayes"**. En dicho trabajo, se realiza la optimización de los clasificadores utilizando algoritmos genéticos tanto con como sin elitismo. En cuanto a su media de porcentaje de clasificación informado en la tabla de resultados, han obtenido utilizando Naive Bayes sin elitismo un máximo de 15.75% y con elitismo un 17.85%, Respecto al clasificador KNN, con k = 3 empleando algoritmos genéticos sin elitismo pudieron obtener un 93.78% y un 94.32% con elitismo.

Como se puede apreciar, los resultados obtenidos difieren de los nuestros debido al enfoque utilizado a la hora de extraer las características del vector x junto con sus parámetros.

Conclusión y trabajos futuros
Los resultados obtenidos probaron que el uso de algoritmos genéticos puede mejorar el desempeño de algoritmos clasificadores utilizados para clasificar arritmias cardiacas a partir de un ECG.
Podemos concluir que como el clasificador GaussianNB es basado en probabilidades es sensible a las clases que se encuentran en mayor proporción (latidos normales), tendiendo a clasificar al resto de latidos dentro de este conjunto. En cambio, KNN no se basa en probabilidades, sino en la distancia entre los latidos en el espacio de características. Esto le permite ser menos sensible a la desproporción en el tamaño de las clases, ya que clasifica un nuevo latido en función de los k vecinos más cercanos, sin importar el tamaño de las clases. Si un latido tiene características similares a otras arritmias, KNN las detectará aunque la clase de los latidos normales sea dominante, esto le permite realizar mejores predicciones en esta aplicación.
Creemos firmemente que el rendimiento y la eficacia de los clasificadores elegidos puede explorarse y mejorarse aún más con la selección de otros conjuntos de pruebas que utilicen características más determinantes, y con el uso de datos ya procesados manualmente por un profesional de la salud y no provenientes de la extracción a partir de las señales de cada ECG. Evitando así las incongruencias entre las anotaciones extraídas por el algoritmo y las detectadas por una persona.

Referencias

Padilla-Navarro, C., González-Reyna, S., Aguilera-González, G., Ortega-Yepez, M., Bombela-Jiménez, S., Rangel-Huerta, M., & Lino-Ramírez, C. (2017). Algoritmos genéticos aplicados a la optimización de características en la clasificación. *Research in Computing Science*, 134, 5-16. https://rcs.cic.ipn.mx/2017_134/Algoritmos%20geneticos%20aplicados%20a%20la%20optimizacion%20de%20caracteristicas%20en%20la%20clasificacion.pdf

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. H. A. (2021). NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53, 1689-1696. https://doi.org/10.3758/s13428-020-01516-y

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. Ch., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

Fortin, F.-A., De Rainville, F.-M., Gardner, M.-A., Parizeau, M., & Gagné, C. (2012). DEAP: Evolutionary algorithms made easy. *Journal of Machine Learning Research*, 13, 2171-2175. https://www.jmlr.org/papers/volume13/fortin12a/fortin12a.pdf

TABLAS

tabla 1

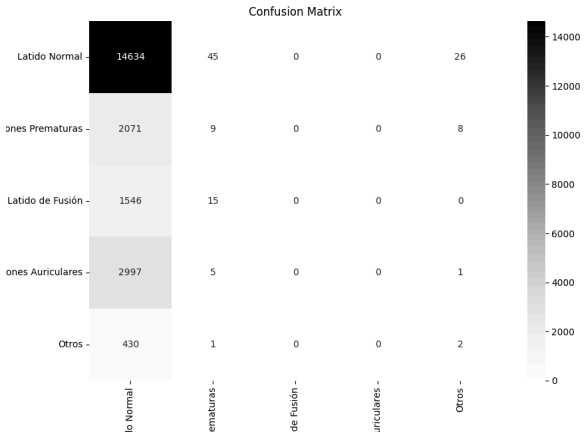

tabla 2

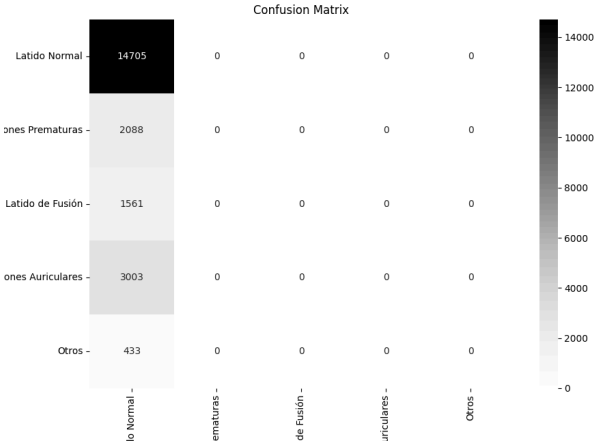

tabla 3

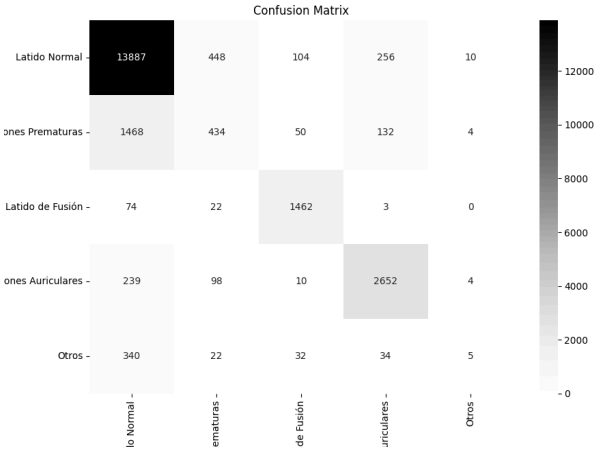

tabla 4