Clasificación de arritmias cardiacas utilizando KNN y Naive Bayes mejorada con algoritmos genéticos (AG) y optimización de cúmulo de partículas (PSO)

Christian Padilla-Navarro[1], Rosario Baltazar-Flores[1], David Cuesta-Frau[2] y Victor Zamudio-Rodríguez[1]

[1] División de Estudios de Posgrado e Investigación, Instituto Tecnológico de León, Av. Tecnológico S/N Col. Industrial Julián de Obregón, 37290, León, México
{jocripana,r.baltazar,vic.zamudio}@ieee.org
http://posgrado.itleon.edu.mx/
[2]Instituto Tecnológico de Informática, Universidad Politécnica de Valencia, Campus Alcoi, Plaza Ferrándiz y Carbonell, 2, 03801 Alcoi, España
{dcuesta}@disca.upv.es
http://www.epsa.upv.es/

Resumen En la presente investigación se busca aumentar el porcentaje de clasificación de las características de una base de datos de arritmias cardiacas aplicando metaheurísticas (Algoritmos Genéticos con y sin Elitismo y Optimización por Cúmulo de Partículas) y diversos clasificadores (KNN con un vecino -1NN-, KNN con tres vecinos -3NN-, KNN con cinco vecinos -5NN- y Naive Bayes), con el fin de realizar una selección de las características principales y reducir el ruido en la base de datos. El aumento en el porcentaje de clasificación en todos los casos fue siempre mayor al 10 %, y la reducción de características llegó a ser en algunos casos hasta del 79 %.

Palabras clave: ECG, arritmia, KNN, Naive Bayes, AG, PSO, reducción de características.

1. Introducción

La función principal de los clasificadores es separar a través de clases los distintos datos con los que contamos. Existen diversas aplicaciones de clasificadores en señales de ECG, que tienen la finalidad de distinguir entre las clases de arritmias existentes. En Nasiri [16] se realizó la clasificación se arritmias provenientes de señales de ECG aplicando Máquinas de Soporte Vectorial (SVM) y Algoritmos Genéticos. Friedman [20] muestra distintas técnicas del Reconocimiento de Patrones, diversos clasificadores y sus variantes. En Badeeh [1] se aplicó el aprendizaje de las máquinas en el diagnóstico de ECG. Ramírez [18] propuso un modelo de clasificación dinámico de arritmias cardiacas mediante aprendizaje de máquina con interfaz a usuario. En Mohamed [22] se presentaron

dos métodos para la clasificación de arritmias multiclase aplicando el Análisis de Componentes Principales (PCA), las Máquinas de Soporte de Vectores Difusos, y el Agrupamiento Desbalanceado. Kallas [19] mostró la clasificación de arritmias multiclase a través de Máquinas de Soporte Vectorial (SVM) combinadas con la extracción de características a través del Análisis de Componentes Principales en señales de ECG. En Thanapatay [8] se propuso un nuevo método de clasificación para ECG aplicando el Análisis de Componentes Principales (PCA) y Máquinas de Soporte Vectorial (SVM). Rabee [4] presentó la clasificación de señales de ECG utilizando Máquinas de Soporte Vectorial y basándose en el análisis de Wavelets de Multiresolución. En Shen [21] se propuso un modelo de clasificación aplicando Máquinas de Soporte Vectorial (SVM) y el Análisis de Componentes Independientes y Zellmer [9] logró la clasificación de señales de ECG basándose en la transformada Wavelet contínua y Máquinas de Soporte Vectorial (SVM).

La combinación de clasificadores con metaheurísticas para la reducción de características en las señales de ECG es una forma eficiente de eliminar datos innecesarios o que generan ruido, ayudando así en la búsqueda de anomalías en los latidos. En Martínez [3] se propone un modelo de clasificación de ECG aplicando sistemas inteligentes para la detección de problemas del corazón aplicando Redes Neuronales y Algoritmos Genéticos. En Melgani [11] se realizó la clasificación de señales de ECG a través de Máquinas de Soporte Vectorial (SVM) aplicando el algoritmo PSO (Particle Swarm Optimization). Mientras que en Daamouche [2] se buscó optimizar la clasificación aplicando el algoritmo de PSO, las Máquinas de Soporte Vectorial (SVM) y realizando la reducción de ruido en las señales a través de Wavelets. En Fira [23] se investigaron los resultados de la clasificación de la compresión de señales de ECG basándose en diversos tipos de matrices de proyección utilizando el algoritmo KNN (K-Nearest Neighbour). En Vaish [7] se investigan las cargas de eficacia y eficiencia computacional de diferentes algoritmos que se utilizan para reconocer el estado emocional a través de señales fisiológicas cardiovasculares, utilizando tablas de decisión, Perceptrón Multicapa, C4.5 y Naive Bayes como un objeto de estudio, la clasificación la realizaron en dos ámbitos: la excitación alta y baja.

La capacidad de identificar automáticamente las arritmias de ECG es importante para el diagnóstico clínico y el tratamiento. En Soman [24] se utilizaron sistemas de aprendizaje automático, Oner, J48 y Naive Bayes para clasificar conjuntos de datos de arritmias obtenidas de ECG médicos. En Gao [6] se describe un sistema para la detección de arritmias cardiacas en señales ECG, basándose en una red neuronal artificial bayesiana (ANN). El clasificador ANN se construye mediante el uso de un modelo de regresión logístico y el algoritmo BackPropagation.

2. Datos de prueba

Para realizar las pruebas utilizamos el conjunto de datos de arritmias del UCI [10] (Center for Machine Learning and Intelligent System). Esta Base de Datos contiene 279 atributos, de los cuales 206 fueron evaluados linealmente y el resto de forma nominal. Dichos datos están basados principalmente en las ondas P, Q, R, S, T y U, y los segmentos que se forman con estas en diversos canales. La base de datos contiene 452 instancias y algunos valores perdidos.

2.1. Función objetivo

Inicialmente se realizó la clasificación de la base de datos con cada uno de los clasificadores propuestos (1NN, 3NN, 5NN y Naive Bayes), obteniendo como función objetivo el porcentaje de clasificación. Posteriormente, se propuso aplicar diversas Metaheurísticas (Algoritmos Genéticos, Genéticos con Elitismo y PSO), siempre sin perder de vista el porcentaje de clasificación como el objetivo principal. En esta investigación no fue considerada la reducción de características como objetivo, pero si como auxiliar en la búsqueda de la mejor clasificación. En todos los casos la selección de características se realizó de manera aleatoria.

3. Algoritmos genéticos (AG)

Formalmente, y de acuerdo a la definición de Goldberg, "Los Algoritmos Genéticos son algoritmos de búsqueda basados en la mecánica de la selección natural y de la genética natural. Combinan la supervivencia con el mejor individuo entre las estructuras de secuencia, con la posibilidad del intercambio de información estructurada, muchas veces de forma aleatoria, para formar un algoritmo de búsqueda que tiene algunas veces la forma de búsqueda que utilizada por los humanos" [14].

3.1. Operadores genéticos utilizados

Selección. El proceso de selección se realizó a través del método Vasconcelos. Para aplicar este método necesitamos ordenar la aptitud de todos los individuos, ascendente o descendentemente, y tomar el individuo más apto y el menos apto.

Cruza. Realizamos la cruza a partir de dos puntos aleatorios. Se toma de la cadena del Padre (mejor individuo) desde la posición 0 hasta el primer punto aleatorio, desde el primer punto aleatorio hasta el segundo punto aleatorio se toma de la cadena de datos de la Madre (peor individuo), y finalmente desde el segundo punto hasta terminar la cadena se obtiene del Padre.

Muta. Se muta un porcentaje de la población, se toma un dato de la cadena de manera aleatoria y se cambia su valor de 0 a 1 o de 1 a 0.

Elitismo. Se toma el mejor individuo y se clona un porcentaje de veces.

En el **Algoritmo 1** se muestra el Algoritmo Genético Básico utilizado en el proceso de clasificación. Se ha demostrado que el elitismo es fundamental en los algoritmos genéticos, en el caso del Algoritmo Genético con Elitismo se realiza la clonación del mejor individuo en un porcentaje de la población.

Algoritmo 1 Algoritmo Genético Simple para la Clasificación de ECG

1: Datos de entrada: tamaño de la población, porcentaje de mutación, llamadas a función, función objetivo -clasificador-(1NN, 3NN, 5NN, Naive Bayes), número de capas.
2: Inicializar una población aleatoria.
3: Evaluar la aptitud de cada uno de los individuos (porcentaje de clasificación).
4: **mientras** (no concluyan las llamadas a función) **hacer**
5: Seleccionar al mejor y al peor individuo (Vasconcelos).
6: Realizar la cruza entre los padres en un bit aleatorio.
7: Mutar un porcentaje de descendientes.
8: Evaluar la aptitud de cada individuo (porcentaje de la clasificación).
9: Mostrar el mejor individuo.
10: **fin mientras**

4. Optimización por cúmulo de partículas (PSO)

PSO es una metaheurística inspirada en la conducta social de de partículas, comunmente aplicada para la solución de problemas de optimización. En 1995 Kennedy y Eberhart [17] desarrollaron el primer algoritmo. El algoritmo puede utilizarse en funciones continuas o binarias.

La actualización de las velocidades de las partículas puede ser vista en la **Ecuación (1)**.

$$v_i = wv_i + \phi_1(GBest_i - x_i) + \phi_2(LBest_i - x_i) \tag{1}$$

Mientras tanto, para realizar la actualización del valor de x_i se emplea una sigmoide, representada con la **Ecuación (2)**.

$$\overrightarrow{Sig(V_{ij})} = \frac{1}{1 + \exp^{-vij}} \tag{2}$$

Algoritmo 2 Algoritmo PSO para la Clasificación de ECG

Datos de entrada: $\phi_1[0,1]$, $\phi_2[0,1]$, tamaño de la población, llamadas a función, función objetivo -clasificador-(1NN, 3NN, 5NN, Naive Bayes), número de capas.
Inicializar una población aleatoria.
Evaluar la aptitud de cada individuo (porcentaje de clasificación).
Tomar al mejor individuo y la mejor aptitud (GBest).
Guardar la primera aptitud de cada individuo (LBest).
Generar velocidades aleatorias $V[0,1]$.
mientras (no concluyan las llamadas a función) **hacer**
  para cada partícula i **hacer**
    para cada miembro de la partícula j **hacer**
      Actualizar velocidades (ver Ecuación 1).
      Actualizar el valor de x_i (ver Ecuación 2).
    fin para
    para cada miembro de la partícula j **hacer**
      Generar un número aleatorio $r_{ij}[0,1]$.
      si $r_{ij} < Sig(V_{ij})$ **entonces**
        $x_{ij} = 0$
      fin si
      si $r_{ij} > Sig(V_{ij})$ **entonces**
        $x_{ij} = 1$
      fin si
    fin para
    Evaluar la aptitud de cada individuo (porcentaje de clasificación).
    Actualizar el LBest.
    si $f(x_i) > f(LBest_i)$ **entonces**
      $LBest_i = x_i$
    fin si
  fin para
  Encontrar el mejor LBest
  si $LBest > GBest$ **entonces**
    $GBest_i = LBest_i$
  fin si
fin mientras

5. Experimentos y resultados

Se realizaron diversas pruebas, todas para 2000 llamadas a función y 10 capas. En el caso de los Algoritmos Genéticos, el porcentaje de mutación utilizado fue del 20 % y 10 % de elitismo, dichos porcentajes fueron propuestos a manera experimental. En el caso de PSO, $\phi_1 = 0.3$, $\phi_2 = 0.5$ y w = 0.7. Las primeras pruebas se hicieron con los clasificadores, 1NN, 3NN, 5NN y Naive Bayes, sin aplicar metaheurísticas. Se consideraron todas las características para tener un parámetro de arranque y poder comparar buscando una mejora. Posteriormente,

se probó con los mismos clasificadores, pero se aplicó el Algoritmo Genético, el Genético con Elitismo y el PSO.

Se puede ver la Media del Porcentaje de Clasificación obtenido (ver la Tabla 1) y la Media del Número de Características empleadas en la clasificación (ver la Tabla 2).

Tabla 1. Media del porcentaje de clasificación.

Algoritmo	**Clasificador**	**Población**	**Media**
Sin algoritmo	1NN	-	52.87 %
Algoritmo Genético	1NN	10	66.27 %
Genético con Elitismo	1NN	10	65.35 %
PSO	1NN	10	64.49 %
Algoritmo Genético	1NN	100	63.41 %
Genético con Elitismo	1NN	100	62.92 %
PSO	1NN	100	64.78 %
Sin algoritmo	3NN	-	57.74 %
Algoritmo Genético	3NN	10	67.71 %
Genético con Elitismo	3NN	10	66.77 %
PSO	3NN	10	67.35 %
Algoritmo Genético	3NN	100	65.71 %
Genético con Elitismo	3NN	100	65.21 %
PSO	3NN	100	67.42 %
Sin Algoritmo	5NN	-	59.29 %
Algoritmo Genético	5NN	10	67.40 %
Genético con Elitismo	5NN	10	64.20 %
PSO	5NN	10	67.28 %
Algoritmo Genético	5NN	100	65.14 %
Genético con Elitismo	5NN	100	65.21 %
PSO	5NN	100	67.35 %
Sin Algoritmo	Naive Bayes	-	61.72 %
Algoritmo Genético	Naive Bayes	10	67.61 %
Genético con Elitismo	Naive Bayes	10	66.41 %
PSO	Naive Bayes	10	68.42 %
Algoritmo Genético	Naive Bayes	100	65.55 %
Genético con Elitismo	Naive Bayes	100	65.27 %
PSO	Naive Bayes	100	67.57 %

Tabla 2. Media del número de características empleadas para la clasificación.

Algoritmo	Clasificador	Población	Media
Sin Algoritmo	1NN	-	279
Algoritmo Genético	1NN	10	138.87
Genético con Elitismo	1NN	10	141.83
PSO	1NN	10	85.26
Algoritmo Genético	1NN	100	139.65
Genético con Elitismo	1NN	100	137.78
PSO	1NN	100	74.14
Sin Algoritmo	3NN	-	279
Algoritmo Genético	3NN	10	126.86
Genético con Elitismo	3NN	10	142.94
PSO	3NN	10	80.04
Algoritmo Genético	3NN	100	147.07
Genético con Elitismo	3NN	100	141.29
PSO	3NN	100	78.07
Sin Algoritmo	5NN	-	279
Algoritmo Genético	5NN	10	134.94
Genético con Elitismo	5NN	10	146.56
PSO	5NN	10	84.54
Algoritmo Genético	5NN	100	141.08
Genético con Elitismo	5NN	100	135.90
PSO	5NN	100	71.07
Sin Algoritmo	Naive Bayes	-	279
Algoritmo Genético	Naive Bayes	10	135.34
Genético con Elitismo	Naive Bayes	10	140.93
PSO	Naive Bayes	10	80.82
Algoritmo Genético	Naive Bayes	100	138.48
Genético con Elitismo	Naive Bayes	100	145.39
PSO	Naive Bayes	100	81.85

5.1. Prueba no paramétrica de los signos de Wilcoxon

Se utilizó la prueba no paramétrica de los signos de Wilcoxon [13] para realizar la comparativa entre los distintos algoritmos, el nivel de significancia empleado fue del 0.1.

Sin algoritmo contra algoritmo genético. Aplicando la prueba de Wilcoxon [13] para realizar la comparativa entre el método Sin Algoritmo (T^+) y Algoritmo Genético (T^-) con diferentes clasificadores, y tomando la media como parámetro de referencia (ver la Tabla 1), obtenemos que:

Dado que T=min(T^-, T^+)=(36, 0)=0 y $T_0 = 5$, podemos concluir que $T \leq T_0$, y podemos aceptar la hipótesis alternativa H_A. Como el objetivo es maximizar el porcentaje de clasificación, los resultados que se encuentran más a la derecha son los del Algoritmo Genético (T^-) y por tanto se obtienen mejores resultados que con el método Sin Algoritmo.

Algoritmo genético contra algoritmo PSO. Aplicando la prueba de Wilcoxon [13] para realizar la comparativa entre el Algoritmo Genético (T^+) y el algoritmo PSO (T^-) con diferentes clasificadores, y tomando la media como parámetro de referencia (ver la Tabla 1), obtenemos que:

Dado que T=min(T^-, T^+)=(10, 11)=10 y $T_0 = 5$, podemos concluir que no se cumple $T \leq T_0$, y no podemos aceptar la hipótesis alternativa H_A. No es posible determinar cual es el algoritmo que se encuentra más a la derecha.

5.2. Algoritmo genético contra algoritmo genético con elitismo

Aplicando la prueba de Wilcoxon [13] para realizar la comparativa entre el Algoritmo Genético (T^+) y el algoritmo Genético con Elitismo (T^-) con diferentes clasificadores, y tomando la media como parámetro de referencia (ver la Tabla 1), obtenemos que:

Dado que T=min(T^-, T^+)=(28, 8)=8 y $T_0 = 5$, podemos concluir que no se cumple $T \leq T_0$, y no podemos aceptar la hipótesis alternativa H_A. No es posible determinar cual es el algoritmo que se encuentra más a la derecha.

Algoritmo genético con elitismo contra algoritmo PSO. Aplicando la prueba de Wilcoxon [13] para realizar la comparativa entre el Algoritmo Genético con Elitismo (T^+) y el algoritmo PSO (T^-) con diferentes clasificadores, y tomando la media como parámetro de referencia (ver la Tabla 1), obtenemos que:

Dado que T=min(T^-, T^+)=(28.5, 7.5)=7.5 y $T_0 = 5$, podemos concluir que no se cumple $T \leq T_0$, y no podemos aceptar la hipótesis alternativa H_A. No es posible determinar cual es el algoritmo que se encuentra más a la derecha.

6. Conclusiones

En todos los casos, el aumento en la clasificación fue mayor al 10 % utilizando una metaheurísitica como auxiliar en la mejora, y la reducción de características llegó hasta el 79 %, lo que conllevaría un ahorro en el tiempo computacional ya que se trabajaría con una cantidad inferior de datos y, por tanto, un avance considerable en caso de ser utilizado en la clasificación en tiempo real en donde es de suma importancia obtener una solución de forma rápida. Además, podemos decir que la reducción de características nos puede llevar a un aumento en el porcentaje de clasificación y por tanto a reducir la cantidad de datos implementados.

Después de haber realizado las pruebas y aplicado la prueba no paramétrica de los signos de Wilcoxon, podemos concluir que es posible implementar una metaheurística como auxiliar en la mejora de la clasificación de señales de ECG, pero no fue posible determinar si PSO, AG o AG con Elitismo nos lleva a los mejores resultados.

Agradecimientos. Agradecemos al proyecto 4573.12-P de la DGEST por su apoyo en esta investigación, y el autor Christian Padilla-Navarro agradece al Consejo Nacional de Ciencia y Tecnología (CONACyT) por la beca otorgada.

Referencias

1. Abdeel-Badeeh M. Salem, Kenneth Revett. Machine learning in electrocardiogram diagnosis. International Multiconference on Computer Science and Information Technology, 978-83-60810-14-9:429-433, 2009.
2. Abdelhamid Daamouche, Latifa Hamami, A wavelet optimization approach for ecg signal classification, Biomedical Signal Processing and Control, 7:342–349, 2012.
3. A.I.Martínez, Rojas, utilización de sistemas inteligentes para la detección de problemas del corazón mediante ecg, 2005.
4. Ayman Rabee, Ecg signal classification using support vector machine based on wavelet multiresolution analysis, The 11th International Conference on Information Sciences, Signal Processing and their Applications: Special Sessions, 978-1-4673-0382-8:1319–1323, 2012.
5. Darwin, C. On the Origin of Species by Means of Natural Selection., John Murray, 1859.
6. Dayong Gao, Michael Madden, Bayesianann classifier for ecg arrhythmia diagnostic system: A comparison study. IJCNN, 2005.
7. Dr. Abhishek Vaish, Performance analysis of machine learning algorithms for emotion state recognition through physiological signal. Global Journal of Computer Science and Technology Neural and Artificial Intelligence, 12, 2012.
8. Dusit Thanapatay, Chaiwat S., Ecg beat classification method for ecg printout with principle components analysis and support vector machines, International Conference on Electronics and Information Engineering (ICEIC), 1:72–75, 2010.
9. Erik Zellmer, Fei Shang. Highly accurate ecg beat classification based on continuous wavelet transformation and multiple support vector machine classifiers, IEEE, 978-1-4244-4134-1, 2009.
10. Frank, A. and Asuncion, A. (2010). UCI Machine Learning Repository `http://archive.ics.uci.edu/ml/datasets/Arrhythmia`. Irvine, CA: University of California, School of Information and Computer Science.
11. Farid Melgani, Classification of electrocardiogram signals with support vector machines and particle swarm optimization, IEEE Transactions On Information Technology In Biomedicine, 12:5, 2008.
12. Fogel L. J. Owens, Artificial Intelligence through Simulated Evolution, 1966.
13. Frank Wilcoxon, Individual comparisons by ranking methods.
14. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning Addison-Wesley Longman Publishing Co., 1989.
15. Holland, Adaptation in Natural and Artificial Systems, 1975.
16. Jalal A. Nasiri, Mahmoud Naghibzadeh, Ecg arrhythmia classification with support vector machines and genetic algorithm, Third UKSim European Symposium on Computer Modeling and Simulation, 978-0-7695-3886-0:187–192, 2009.
17. Kennedy J., Particle Swarm Optimization Proceedings of the 1995, IEEE international conference on neural networks, (ICNN'95), 1995, 4, 1942-1948.
18. López, Modelo de clasificación dinámico de arritmias cardiacas mediante aprendizaje de máquina con interfaz de usuario. Ciencia y Tecnología Neogranadina, 16/002:56–95, 2006.

19. Maya Kallas, Clovis Francis, Multi-class svm classification combined with kernel pca feature extraction of ecg signals. International Conference on Telecomunications, 978-1-4673-0747-5, 2012.
20. Menahem Friedman, Introduction to Pattern Recognition statical, structural, neural and fuzzy logic approaches, Imperial College Press, 1999.
21. Mi Shen, Liping Wang. Multi-lead ecg classification based on independent component analysis and support vector machine, 3rd International Conference on Biomedical Engineering and Informatics (BMEI), 978-1-4244-6498-2:960–964, 2010.
22. Mohamed Cherif Nait-Hamoud, Two novel methods for multiclass ecg arrhythmias classification based on pca, fuzzy support vector machine and unbalanced clustering, IEEE, 978-1-4244-8611-3:140-145, 2010.
23. Monica Fira, Liviu Goras, On the projection matrices influence in the classification of compressed sensed ecg signals, International Journal of Advanced Computer Science and Applications, 3, 2012.
24. Thara Soman, Classification of arrhythmia using machine learning techniques, ICOSSE, 2005.