Traballo Fin de Máster

Aplicación de técnicas de aprendizaxe profunda para o cribado de fibrilación auricular mediante a clasificación automática de rexistros de ECG

María Bugallo Porto

Máster en Técnicas Estatísticas

Curso 2021-2022

Proposta de Traballo Fin de Máster

Título en galego: Aplicación de técnicas de aprendizaxe profunda para o cribado de fibrilación auricular mediante a clasificación automática de rexistros de ECG

Título en castellano: Aplicación de técnicas de aprendizaje profundo para el cribado de fibrilación auricular mediante la clasificación automática de registros de ECG

English title: Deep learning for screening of atrial fibrillation on ECG recordings

Modalidade: Modalidade A

Autora: María Bugallo Porto, Universidade de Santiago de Compostela

Directores: Paulo Félix Lamas, Universidade de Santiago de Compostela; Balbina Virginia Casas Méndez, Universidade de Santiago de Compostela

Breve resumo do traballo:

i. Introdución do problema de clasificación de rexistros de electrocardiograma (ECG). Importancia clínica e revisión das aportacións da literatura previas ao traballo. Motivación, contexto e procedencia dos datos empregados.
ii. Marco teórico da memoria. Exposición das técnicas de aprendizaxe supervisada aplicadas, entre as que destacan as redes neuronais LSTM, como mellora das redes neuronais recorrentes, as redes de atención e o clasificador global XGBoost. Aplicación das SVM. Estudo das métricas utilizadas como medidas de calidade da clasificación.
iii. Presentación e construción da arquitectura do clasificador proposto. Adestramento e validación da rede e optimización dos hiperparámetros. Selección de variables globais e secuenciais. Interpretabilidade dos resultados obtidos.

Recomendacións: Coñecementos básicos de intelixencia artificial e, en concreto, de aprendizaxe estatística (con especial atención aos modelos de *boosting*, ás redes neuronais artificiais e ás máquinas de soporte vectorial), así como de series de tempo e estimación non paramétrica da densidade. Manexo de `Python` (2022) e `R Core Team` (2022).

Don Paulo Félix Lamas, investigador adscrito ao Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS) e profesor titular da Universidade de Santiago de Compostela, e dona Balbina Virginia Casas Méndez, profesora titular da Universidade de Santiago de Compostela, informan que o Traballo Fin de Máster titulado

Aplicación de técnicas de aprendizaxe profunda para o cribado de fibrilación auricular mediante a clasificación automática de rexistros de ECG

realizouse baixo a súa dirección por dona María Bugallo Porto para o Máster en Técnicas Estatísticas. Estimando que o traballo está rematado, dan a súa conformidade para a súa presentación e defensa ante un tribunal.

En Santiago de Compostela, a 31 de xaneiro de 2022.

O director:

FELIX LAMAS PAULO MANUEL - 36100889C
Firmado digitalmente por FELIX LAMAS PAULO MANUEL - 36100889C
Fecha: 2022.01.31 09:40:32 +01'00'

Don Paulo Félix Lamas

A directora:

CASAS MENDEZ BALBINA VIRGINIA - 09367287P
Firmado digitalmente por CASAS MENDEZ BALBINA VIRGINIA - 09367287P
Fecha: 2022.01.30 16:22:54 +01'00'

Dona Balbina Virginia Casas Méndez

A autora:

BUGALLO PORTO, MARIA (FIRMA)
Firmado digitalmente por BUGALLO PORTO, MARIA (FIRMA)
Fecha: 2022.01.30 15:27:42 +01'00'

Dona María Bugallo Porto

Agradecementos

Á Universidade de Santiago de Compostela e, en particular, ao Vicerreitorado de Investigación e Innovación, polo seu firme compromiso coa investigación e polas distintas axudas destinadas a estudantes de máster, que dotaron de servizos e recursos a realización deste traballo.

Ao Servei d'Estadística Aplicada e ao Institut d'Estadística de Catalunya, por recoñecer esta obra co Premio Almirall ao mellor traballo de bioestatística, no eido do XIX Concurs Student d'Estadística Aplicada, baixo o título *"Redes neuronales basadas en el paradigma de atención: Cribado de fibrilación auricular mediante la clasificación automática de registros de ECG"*.

A todo o persoal do Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), por axudarme en canto precisei e polo inmellorable trato recibido ao longo deste último ano. Desistencionadamente, conseguiron que a miña elección anticipada de realizar o TFM no centro fora a mellor que puiden tomar. Grazas por guiarme no mundo da investigación e ensinarme a crecente importancia da estatística e da intelixencia artificial na actualidade.

Debo agradecer de maneira especial ao meu director, Paulo Félix, o seu trato próximo e a súa paciencia, apoio e dedicación, que fixeron posible esta memoria. Sempre estivo disposto a axudarme e aconsellarme ante calquera problema, cunha actitude indubidablemente positiva que impulsou as miñas ganas de continuar. Grazas tamén ao seu compañeiro de traballo, Jesús Presedo, pola imprescindible colaboración e consellos, moi importantes na interpretación clínica dos resultados.

Á miña directora, Balbina Casas, polo inmensurable esforzo na elaboración e corrección do documento, así como pola súa participación, cercanía e entrega.

En último lugar, pero non menos importante, agradecer á miña familia, a Eduardo e aos meus amigos o seu apoio incondicional e comprensión. Grazas ao seu afecto e consellos, tanto académicos como persoais, logrei rematar con éxito e felicidade os meus estudos universitarios.

Índice xeral

Resumo

Resumo en galego

Neste traballo abórdase a clasificación automática de rexistros de electrocardiograma (ECG), utilizando técnicas de aprendizaxe estatística, para proporcionar un método interpretable destinado ao cribado de fibrilación auricular, propoñendo mecanismos capaces de detectar outras patoloxías e recoller información para o posterior diagnóstico dunha posible enfermidade cardíaca. A novidade da proposta radica en presentar un modelo de clasificación de secuencias baseado no paradigma de atención, previamente aplicado con éxito en problemas de tradución e transcrición lingüística, que destine a información latexo a latexo á creación dunha ferramenta eficaz de cribado da poboación. Esta reforzarase cun clasificador global que empregará coñecemento de toda a secuencia, é dicir, do rexistro completo, para a correcta caracterización das arritmias cardíacas. Destacar que este tema xa se tratou en moitas ocasións e son múltiples as aportacións existentes na literatura.

O noso enfoque non só radica en mellorar certas medidas de calidade da clasificación, senón que se pretende entender as relacións que subxacen detrás das diferentes características que permiten dotar de interpretabilidade á clasificación, principal carencia dos modelos baseados en redes neuronais artificiais e, en xeral, da aprendizaxe profunda. O obxectivo final é a aplicabilidade do modelo na construción dun clasificador con taxas de especificidade e sensibilidade elevadas, en canto á correcta detección de fibrilación auricular, e que sexa auto–explicable no sentido de indicar que propiedades do electrocardiograma son as que determinan o resultado da clasificación.

English abstract

In this document we will treat the automatic classification of electrocardiogram (ECG) recordings, using statistical learning techniques, to provide an interpretable method for atrial fibrillation identification, proposing mechanisms that are able to detect other pathologies and collect information for the subsequent diagnosis of a possible heart disease. The novelty of the proposal is to present a sequence classification model based on the attention paradigm, previously used successfully in translation and transcription problems, that allocates beat-to-beat information to create an effective identification tool. This will be reinforced with a global classifier that uses knowledge of the whole sequence, that is, the complete record, for the correct characterization of cardiac arrhythmias. It should be noticed that this topic has already been dealt on many occasions and there are many contributions to the literature.

Our approach is not only to improve certain quality measures of the classification, but also to understand the relationships behind the different characteristics that allow us to give interpretability to the classification, the main problem of the models based on artificial neural networks and, in general, of the deep learning. The ultimate goal is the applicability of the model in the construction of a classifier with high rates of specificity and sensitivity, in terms of the correct detection of atrial fibrillation, and that is self–explanatory in the sense of indicating which characteristics of the electrocardiogram are those that determine the outcome of the classification.

Prefacio

Na actualidade as enfermidades cardiovasculares son unha das principais causas de morte e ocupación hospitalaria, incrementando gravemente a saturación dos sistemas nacionais de saúde. Por este motivo, estase a promover a nivel global a investigación en ferramentas que permitan un diagnóstico automático de patoloxías cardíacas, comunmente baseadas no procesado dixital de sinais clínicas, tales como os rexistros de electrocardiograma, para facilitar a tarefa médica de diagnóstico e incrementar a autonomía dos pacientes na realización do proceso. En concreto, destaca o uso da intelixencia artificial e, en particular, dos modelos de aprendizaxe estatística, como mecanismos capaces de explorar e modelar a información relevante dos datos.

En relación con isto, a fibrilación auricular (FA) é unha arritmia cardíaca que constitúe un problema de saúde pública extremadamente custoso, afectando a entre o 1 % e o 2 % do total da poboación, sendo unha doenza presente en máis do 5 % da poboación anciá, aumentando progresivamente coa idade, ata un 15 % en suxeitos de 80 anos (Lip et al., 2016). Tendo en conta a tendencia demográfica actual, prevese que a incidencia e repercusión da FA se incremente preocupantemente no futuro. En termos médicos, esta doenza caracterízase por presentar, en condicións normais de repouso, un ritmo cardíaco irregular e a miúdo moi rápido, distinguido por unha distancia temporal entre latexos que non responde a ningún patrón rítmico previsible (distancia arrítmica). En concreto, o seu nome débese á causa física que a xera: o corazón fibrila, como se dun tremor se tratase, limitando en boa medida a súa capacidade funcional. Ademais, esta cardiopatía está asociada cun significativo aumento da mortalidade e do risco de padecer graves complicacións, como o ictus, os accidentes cerebrovasculares, a insuficiencia cardíaca ou o deterioro cognitivo (Savelieva and Camm, 2008).

A proba diagnóstica típica na detección de patoloxías cardíacas é o electrocardiograma (ECG, do alemá *elektrokardiogramm*), instrumento clave no diagnóstico de FA, e que se resume nunha representación visual da actividade eléctrica do corazón, medida na superficie corporal, a través da voltaxe asociada a cada latexo. Para realizar un ECG estándar utilízanse 12 derivacións, é dicir, 12 tomas de diferencia de voltaxe entre dous electrodos, cada un deles situado na parte do corpo que lle corresponde, obtendo 12 "películas" diferentes da actividade eléctrica cardíaca dende varios ángulos. Esta proba, habitualmente realizada a nivel hospitalario, require colocar electrodos no peito, nocellos e pulsos do paciente, e ten unha duración duns 3-4 minutos. Sen embargo, actualmente é cada vez máis frecuente o uso de dispositivos portátiles (*wearables*), resultando crucial obter criterios automáticos que permitan diagnosticar unha patoloxía en base á interpretación dun ECG, optimizar a súa eficiencia e a autonomía do paciente no proceso. Ademais, os dispositivos portátiles reducen o custe sanitario, en comparación coas técnicas habituais, pois non requiren unha preparación previa nin a presenza de persoal sanitario especializado, e a súa demanda computacional é moi baixa. Adicionalmente, a inmensa maioría utilizan un rexistro de ECG dunha soa derivación, o que facilita o seu manexo e adquisición.

A fibrilación auricular pode ser unha doenza asintomática aínda que, frecuentemente, provoca latexos rápidos e fortes (palpitacións), falta de aire ou debilidade. Así mesmo, os seus episodios son de duración variable, sendo nalgúns pacientes esporádicos (FA paroxística) e noutros duradeiros (FA persistente) ou sen tratamento que poida interrompelos ou evitalos (FA crónica), o que dificulta o

cribado. Por outra parte, a interpretación dun rexistro de ECG dunha soa derivación é un problema difícil en si mesmo, presentando unha baixa concordancia ao comparar a opinión de varios expertos cardiólogos, medida en base ao coeficiente Kappa de Cohen (López de Ullibarri and Pita, 1999). Isto tradúcese nunha distinción autónoma complicada das arritmias cardíacas, entre as que se atopa a FA.

En vista da inestimable relevancia da detección precoz de FA e da notable dificultade desta tarefa, constitúe un interesante desafío científico presentar un método automático para o seu cribado, que permita unha correcta clasificación dun rexistro, o que propicia a temática do presente traballo.

Ante esta problemática, e co obxectivo de fomentar o estudo de algoritmos de clasificación adicados á detección de arritmias cardíacas –con especial interese no cribado de FA– a sociedade estadounidense *Physionet* presentou o reto *Computing in Cardiology* (CinC) *Challenge* 2017 (Clifford et al., 2017), titulado *"AF classification from a short single lead ECG recording"*. Este desafío focalízase na discriminación de rexistros de ECG curtos (de 9 a 61 segundos de duración) dunha única derivación (en terminoloxía médica, a primeira derivación)[1] en 4 clases diferentes –coñecidas *a priori*–, codificadas como **N** (ritmo normal), **FA** (fibrilación auricular), **O** (outros ritmos) e ∽ (ruído). Con este fin, puxéronse a disposición da competición 12186 rexistros –doados pola empresa americana AliveCor, especializada na implementación de modelos de intelixencia artificial con finalidade médica en dispositivos móbiles– divididos en dous conxuntos: 8528 nun conxunto público de adestramento (de etiqueta dispoñible) e 3658 nun conxunto privado de test (de etiqueta oculta/secreta). No momento da elaboración e presentación deste documento, o conxunto de test continúa privado e non se pode usar.

A medida de puntuación do reto foi a métrica $F1$ de calidade da clasificación, definida no caso binario como a media harmónica entre a precisión (*precision*) e a sensibilidade (*recall*), e facilmente xeralizable a varias clases. Como a etiqueta de ruído se asocia a rexistros de moi baixa calidade, e non a un ritmo cardíaco en si mesmo, resulta lóxico excluíla da análise do rendemento dos modelos, polo que a medida final a considerar se define como a media aritmética das medidas $F1$ das restantes clases:

$$F1 = \frac{F_{1N} + F_{1FA} + F_{1O}}{3}, \tag{1}$$

onde se definen as medidas $F1$ de cada clase como

$$F_{1N} = \frac{2Nn}{N_T + n_T}, \quad F_{1FA} = \frac{2Aa}{A_T + a_T} \quad \text{e} \quad F_{1O} = \frac{2Oo}{O_T + o_T},$$

sendo Ll o total de rexistros de certa clase correctamente clasificados e L_T e l_T o total de rexistros etiquetados ou clasificados como de certa clase, respectivamente, con $(L, l) \in \{(N, n), (A, a), (O, o)\}$.

A importancia deste desafío na elaboración de modelos destinados ao cribado de FA, con aplicación directa no uso clínico, reside na riqueza da fonte de datos. Anteriormente, realizáronse estudos previos destinados ao cribado de FA, pero existen razóns que nos permiten afirmar que a súa aplicabilidade é moi limitada: ou ben só consideraban pacientes normais e pacientes enfermos de FA, con rexistros demasiado preparados, desbotando a presenza doutras patoloxías; ou ben os datos estaban coidadosamente seleccionados, incorporando só rexistros pouco ruidosos e exemplos das clases moi claros; ou ben o tamaño mostral era insuficiente ou as derivacións eran 12, o que non permite avanzar na autonomía dos pacientes no proceso; etc. Sen embargo, no eido do *Physionet*/CinC *Challenge* 2017, os rexistros son curtos e dunha soa derivación, contando coa presenza de múltiples arritmias, algunhas delas semellantes á FA. Ademais, as referencias caracterízanse por presentar fragmentos altamente influenciados polo ruído e por artefactos mostrais, imitando a súa frecuencia de aparición na realidade.

[1] A primeira derivación, ou derivación LA-RA, é unha derivación bipolar que mide a diferencia de voltaxe entre os electrodos LA (*left arm*) e RA (*right arm*), situados nas monecas esquerda e dereita do paciente.

Como consecuencia da dispoñibilidade e validez desta información, o adestramento e validación do modelo proposto no presente traballo utiliza como fonte de datos o conxunto público de rexistros e etiquetas deste reto. A métrica $F1$ permitiranos comparar os nosos resultados coas mellores aportacións presentadas á competición e extraer conclusións acerca do rendemento do clasificador proposto e das dificultades do cribado de FA –mediante a clasificación automática de rexistros de ECG curtos e dunha soa derivación– así como confirmar as complicacións asociadas á correcta detección doutros ritmos anómalos e á discrepancia presente entre os cardiólogos anotadores. Destacar que o esmero nas melloras realizadas centrarase no incremento da interpretabilidade dos resultados, co obxectivo de entender os patróns asimilados, que determinan a discriminación das clases. Como última meta, preténdese achar e discutir as debilidades e fortalezas do clasificador final.

Unha vez presentado o propósito do traballo, así como o problema médico que o motiva, o resto do documento ocúpase do seu estudo mediante ferramentas de aprendizaxe estatística e de validación de modelos. Cabe mencionar que o problema de clasificación multiclase e o cribado de FA se tratarán dende o Capítulo 1 e, por clarificar a presentación, a nosa elección foi introducir os diferentes elementos matemáticos e estatísticos no momento no que se precisen, discutindo as súas vantaxes e limitacións.

En canto á estrutura xeral do documento, divídese en catro capítulos e axudarémonos de tres apéndices, con resultados e explicacións adicionais, claves para conseguir unha memoria autocontida e unha investigación reproducible. No Capítulo 1 preséntanse, de maneira xeral, os elementos de aprendizaxe estatística e recoñecemento de patróns que se usarán e detallarán ao longo da memoria. Tamén se introduce o traballo principal no que se apoian os progresos acadados e as ferramentas informáticas utilizadas, con especial atención á metodoloxía propia do *framework Construe* (Teijeiro and Félix, 2018). No Capítulo 2 especifícase o clasificador proposto, deténdonos na xustificación da súa estrutura e na explicación e formalización matemática dos modelos involucrados, incluíndo o proceso de selección de variables e o adestramento e optimización dos hiperparámetros. Adicionalmente, pormenorízase na interpretabilidade da solución, dende a súa relación coa formulación matemática á inclusión de exemplos gráficos ilustrativos. O Capítulo 3 cubre a análise dos resultados para, finalmente, completar o traballo no Capítulo 4, coa exposición das conclusións experimentais e os intentos de mellora realizados, indicando as liñas de investigación futuras. Por outra banda, o Apéndice A contén os resultados principais do *Physionet*/CinC *Challenge* 2017, con especial atención á aportación dun dos gañadores. Seguidamente, o Apéndice B destínase á descrición das variables de entrada dos modelos, incluíndo o resultado da súa selección e o seu preprocesado. Finalmente, no Apéndice C precísase o adestramento e optimización dos hiperparámetros dos modelos, engadindo os seus valores finais.

Capítulo 1

Preliminares e material empregado

Neste capítulo introdúcese o marco de estudo onde se localiza o problema a resolver, así como a exposición xeral deste último e os avances previos ao presente traballo –reproducidos aquí con éxito–, con especial atención ao estudo de investigación realizado por Teijeiro et al. (2018b). Para este cometido, o capítulo divídese en seis seccións e a súa organización explícase a continuación.

Primeiramente, a Sección 1.1 abarca a introdución dos elementos de aprendizaxe estatística e recoñecemento de patróns que se usarán ao longo da memoria. Dende unha perspectiva xeral, defínense conceptos tales como a aprendizaxe profunda e aclárase a principal finalidade do noso enfoque, a interpretabilidade dos resultados de clasificación. Na Sección 1.2 preséntase a aportación principal na que se apoia a nosa proposta (sobre a que se profundizará no Capítulo 2), introducindo o *framework* *Construe*, explicado na penúltima sección. A continuación, a Sección 1.3 recolle detalles relativos á fonte de datos, que complementarán a descrición realizada no Prefacio, e na Sección 1.4 preséntase a estrutura básica dun latexo e as súas compoñentes morfolóxicas principais, esenciais para a caracterización e comprensión das variables globais e secuenciais utilizadas como entrada dos nosos modelos. Finalmente, a Sección 1.5 destínase á presentación e descrición de *Construe* e na Sección 1.6 recóllese unha especificación das ferramentas informáticas utilizadas, con especial interese nas linguaxes de programación `R Core Team` (2022) e `Python` (2022) e nas súas librarías máis destacables.

1.1. Aprendizaxe estatística e recoñecemento de patróns

Neste traballo faise un uso intensivo de técnicas propias da aprendizaxe estatística (*statistical learning*), importante rama da intelixencia artificial (*artificial intelligence*, AI), cuxa finalidade é o desenvolvemento de mecanismos destinados á instrución das computadoras, no sentido de que o seu desempeño mellore coa experiencia (co consumo de datos), para posteriormente tomar decisións cunha intervención humana mínima. Desta forma, partindo dunha mostra (conxunto de adestramento), apréndese a recoñecer como as características dos datos permiten a súa identificación coas distintas categorías obxecto da aprendizaxe, ensinando ao modelo cara unha posterior aplicación a datos nunca vistos (conxunto de test). Neste sentido, os clasificadores presentados serán modelos de aprendizaxe supervisada (*supervised learning*) porque os datos de adestramento están previamente etiquetados e procúrase a súa clasificación en catro clases prefixadas: **N**, **FA**, **O** e ∽. De feito, o nome destes modelos débese á dispoñibilidade dun conxunto de etiquetas de saída, que guía o proceso de aprendizaxe.
Contrariamente, nos problemas de aprendizaxe non supervisada (*unsupervised learning*) non se dispón dunha etiquetaxe previa dos datos, senón que se debe aprender a partir dun conxunto de variables de entrada. Deste xeito, é necesario establecer como se organizan e agrupan as observacións en función do comportamento dunha mostra de adestramento.

Na aprendizaxe estatística trátase de comprender –na medida do posible– o proceso subxacente de

xeración de datos e se estes son representativos da poboación (é dicir, se non están nesgados). Como comentamos no Prefacio, a maioría de estudos anteriores ao *Physionet*/CinC *Challenge* 2017 estaban nesgados porque, ou ben os datos non representan fielmente á diversidade da poboación de rexistros de ECG –por só considerar exemplos de **N** e de **FA**– ou ben se escollían exemplos demasiado claros.

Directamente relacionado coa aprendizaxe estatística, o recoñecemento de patróns (*pattern recognition*) encárgase de construír mecanismos capaces de extraer información relevante e pautas clave das observacións dunha mostra. Deste xeito, ocúpase da identificación de regularidades nos datos, co propósito de impoñer un conxunto de relacións de identidade (clasificación, agrupamento, asociación, etc.) ou dependencia (regresión) entre os seus elementos. Neste caso, a finalidade é determinar que características do ECG son propias dun paciente enfermo de FA e distinguilas das de calquera outro incidente patolóxico, á par que caracterizar cales pertencen a un individuo san e modelar que descubrimentos son propios dun rexistro ruidoso. Esta tarefa realizaraa o *framework Construe* (Teijeiro and Félix, 2018), ferramenta clave para nós no recoñecemento de patróns no ECG, que enriquecerá aos modelos de clasificación con dous conxuntos de variables (un deles destinado á análise global do rexistro e outro á secuencial) obtidas automaticamente a partir da descrición rítmica e morfolóxica dos rexistros de ECG, intentando imitar os criterios médicos de diagnóstico, baixo o etiquetado do *PhysioNet*/CinC *Challenge* 2017. A súa explicación máis detallada preséntase na Sección 1.5.

Resumidamente, a entrada de *Construe* é un rexistro de ECG que, por intuición, vén dado por unha serie temporal en tempo discreto. Como resultado, obtense unha descrición estatística dos rexistros por medio das variables proporcionadas polo *framework*, resultando axeitado introducir a súa esencia e motivación. Por unha banda, un conxunto de variables secuenciais extráese de cada latexo, modelando o seu comportamento individual como integrantes da cadea completa, permitindo detectar latexos ectópicos e incidentes electrocardiográficos patolóxicos de curta duración (de un ou dous latexos). Por outra banda, e para mellorar a súa caracterización, extráese un segundo conxunto de variables, referido a medidas resumo da secuencia completa, permitindo detectar patoloxías prolongadas no tempo ou tal que sexa a súa frecuencia o que orixine o carácter anómalo da referencia. Seguidamente, e como se precisará no Capítulo 2, cada conxunto de variables conforma a entrada dun clasificador individual: un modelo de *boosting* e unha rede neuronal adaptada á natureza do problema combínanse mediante unha técnica de *ensemble* para proporcionar unha resposta fundamentada nos valores das variables e na súa asociación coa etiquetaxe da fonte de datos.

Deste xeito, na presente memoria combínanse nocións de aprendizaxe estatística, centrándonos nun modelo de *boosting* e nas redes neuronais artificiais, para modelar información extraída para a clasificación automática de rexistros de ECG, destinada á detección de arritmias cardíacas e, en particular, ao cribado de FA. Por este motivo, adicaremos especial atención á aprendizaxe profunda (*deep learning*), subcategoría da aprendizaxe estatística relativa ao uso de redes neuronais na localización de patróns en ambientes caracterizados por certa aleatoriedade. En especial, a nosa proposta ofrece unha primeira aproximación para solventar as dificultades do cribado de FA, presentando por primeira vez na literatura un modelo de rede neuronal interpretable para abordar a detección de arritmias cardíacas en base á análise dun rexistro de ECG curto dunha soa derivación, realizado cun dispositivo móbil a nivel ambulatorio. A finalidade é buscar una proba médica de baixo custo, coa opción dunha realización autónoma e sinxela por parte do paciente, sen persoal sanitario especializado, procedendo dende un enfoque propio: enfatízase a interpretabilidade dos criterios de clasificación e, desta maneira, ofrécese unha contribución diferente, que permite melloras en relación aos resultados previos da literatura.

Tipicamente, o conxunto de variables de entrada denotarase coa letra maiúscula $\overrightarrow{X}$, que vai ser un vector aleatorio con D compoñentes (atributos): considéranse D variables de entrada, representadas mediante X_j, para $j = 1, ..., D$. A observación i-ésima de $\overrightarrow{X}$ denotarase por $\overrightarrow{x_i}$, tal que

$$\overrightarrow{x_i} = (x_{i1}, ..., x_{iD}) \in \mathbb{R}^D, \text{ para } i = 1, ..., n \text{ observacións.}$$

Neste caso, unha mostra de tamaño n dun vector de D variables constitúe unha matriz $n \times D$ e denotaremos ao espazo de tales matrices coa grafía $\mathcal{M}_{n \times D}$. Ademais, a anterior matriz almacena por columnas cada variable coas súas n observacións e, por filas, cada dato mostral. Así mesmo, as matrices vanse representar en maiúsculas e reservaremos letras tales como i, j, k, m e t para referirnos a subíndices e superíndices, elixindo t para a dimensión temporal e restrinxindo o seu uso só aos subíndices[1].

A variable resposta, tanto no caso continuo[2] como no caso discreto, denotarase por Y. No primeiro deles, é claro que $Y \in \mathbb{R}$ (regresión) e no segundo, traballaremos con variables cualitativas, codificadas para a súa representación numérica, con exemplos de $Y \in \{0, 1\}$, $Y \in \{-1, 1\}$ e $Y \in \{1, ..., K\}$, para $K > 1$ clases (clasificación binaria e clasificación multiclase). Por analoxía coas variables de entrada, para cada $i \in \{1, ..., n\}$, a i-ésima observación de Y denotarase por y_i, indistintamente.

1.2. Aportación destacada ao *Physionet*/CinC *Challenge* 2017

No reto participaron 75 equipos independentes, con variedade de modelos tradicionais e novos, dende as básicas árbores de decisión a modelos de aprendizaxe profunda, como as redes neuronais artificiais e as súas versións recorrentes e convolutivas. Finalmente, e como se explica no Apéndice A, catro equipos gañaron o desafío, cunha medida $F1$ promediada nas clases do conxunto $\{\mathbf{N}, \mathbf{FA}, \mathbf{O}\}$ igual a 0.83. Entre todas as aportacións, debemos destacar o estudo de investigación exposto en Teijeiro et al. (2018b), posicionado entre os catro con mellor desempeño –e que constitúe o piar fundamental deste traballo– que utiliza redes neuronais *long short-term memory* (LSTM) para a clasificación secuencial, xunto cunha mellora das árbores de decisión, o XGBoost, para a clasificación global.

O éxito da súa contribución radica na aplicación dun método de *ensemble* (ou método de *stacking*), combinando un clasificador global e un secuencial, para mellorar a xeralización e rendemento individual dos modelos de nivel inferior. Concretamente, a proposta acadou unha medida $F1$ de 0.83 na fase xeral (sobre o conxunto de test oculto) –resultando unha das gañadoras do desafío–. Na fase de seguimento posterior, e cunha significativa simplificación do modelo, mellorouse a 0.85. Como se pode ver na Figura 1.1, este último valor ocasiona que a proposta mencionada destaque sobre todas as demais e constitúe a puntuación máis alta acadada ata o de agora para o conxunto de test oculto.

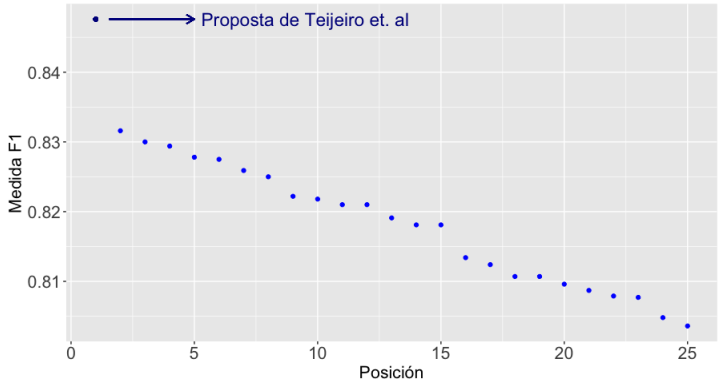


Figura 1.1: Desempeño dos 25 mellores algoritmos no seguimento do *Physionet*/CinC *Challenge* 2017.

No noso caso, e seguindo a idea anterior, tamén se aplicará un método de *ensemble*, combinando unha clasificación global e unha secuencial. No modelo secuencial, ímonos apoiar na construción dunha

[1] O uso do superíndice t resérvase para denotar a versión trasposta dunha matriz ou vector.

[2] Na formalización do clasificador global realizada na Sección 2.1, requírese introducir un algoritmo presentado para problemas de regresión (resposta continua) e que será fundamental para entender a formulación do modelo global.

arquitectura de rede neuronal recorrente adaptada á natureza do problema, e que mellorará os resultados acadados no estudo anterior. Como mencionamos previamente, e para dotar de coñecemento aos clasificadores, dous conxuntos de características foron extraídos dos sinais de ECG –as características globais e as secuenciais– e usados como variables entrada de dous modelos de clasificación, con distinta natureza e dedicación. En concreto, coinciden coas variables utilizadas en Teijeiro et al. (2018b), o que motiva a explicación do proceso de preprocesado do sinal e recoñecemento de patróns realizado nese estudo. Para ambas tarefas, utilízase o *framework Construe*, unha ferramenta baseada no razoamento abdutivo para a interpretación de series de tempo, xunto cun algoritmo para a clasificación automática de latexos en sinais de ECG (Teijeiro et al., 2018a). Ambos se describen na Sección 1.5, logo de comentar unha serie de detalles relativos á base de datos, indicando a versión da etiquetaxe utilizada e fixando notación clave en explicacións futuras.

1.3. Fonte de datos

Tal e como comentamos no Prefacio, a base de datos usada no adestramento e validación do modelo proposto correspóndese co conxunto público do *Physionet*/CinC *Challenge* 2017, contando cun total de 8528 rexistros. Adicionalmente, a versión do etiquetado utilizada non foi a orixinal, senón a versión V3, posterior a tres procesos de reetiquetado, labor realizado polos organizadores do desafío con motivo da gran discrepancia cometida entre os anotadores, e que utilizou resultados do propio reto para mellorar a calidade da información (Clifford et al., 2017). A idea desta última versión foi seleccionar os rexistros onde os 10 primeiros modelos discrepaban máis, obtendo 1129 referencias tal que a súa clasificación semellaba ser *demasiado* variable entre os mellores clasificadores, co obxectivo de valorar a lexitimidade da súa clase. Nesta tarefa, oito cardiólogos revisaron as 1129 etiquetas e, finalmente, moitas foron modificadas para dar lugar á versión V3.

En cando á definición dos integrantes da base de datos, debemos mencionar que un rexistro de ECG constitúe unha medición da actividade eléctrica do corazón ao longo do tempo: rexistramos a actividade eléctrica cardíaca durante varios segundos (entre 9 e 61 segundos) dende a primeira derivación. Deste xeito, cada rexistro é unha serie de tempo[3] independente e a nosa tarefa é a súa correcta clasificación. Clarificar que as observacións se tomaron cunha frecuencia de 300 Hz, de modo que nun rexistro de 30 segundos contamos con 9000 datos, que representamos mediante a súa gráfica secuencial: dada unha serie de tempo, é fundamental representar cada dato fronte ao instante de observación e logo unir cada un dos puntos mediante segmentos. Na Figura 1.2 inclúense catro series de tempo correspondentes aos rexistros A00140 (**N**), A00102 (**FA**), A00741 (**O**) e A00022 (∽), representando o tempo (en segundos, s) fronte á voltaxe (en voltios, V).

Unha primeira análise visual permítenos intuír que cada unha delas ten unha aparencia claramente distinta. Máis adiante veremos que será esta estrutura, tanto de forma global como latexo a latexo, a que capacite a súa clasificación, buscando patróns e feitos excluíntes/comúns a cada unha das clases.

Fixando un pouco de notación, sexa n o tamaño mostral (cantidade de rexistros de ECG dispoñibles no conxunto público de adestramento do *Physionet*/CinC *Challenge* 2017) e K o número de clases consideradas. Teoricamente, $n = 8528$ referencias e $K = 4$ clases. Sen embargo, o tamaño mostral tívose que reducir en 2 unidades por culpa dunha serie de problemas de procesado detectados en dúas referencias: os rexistros A02133 (**O**) e A07460 (**O**) non puideron utilizarse por mor dun erro interno

[3]Dado un proceso estocástico (*stochastic process*, SP) en tempo discreto (por exemplo, en $\mathbb{Z}$) e con espazo de estados continuo (por exemplo, $\mathbb{R}$), $\{X_t\}_{t\in\mathbb{Z}}$, supoñamos que se observa un fragmento da súa traxectoria, $(x_1, ..., x_T)$, conformando este conxunto de datos unha mostra de T valores dependentes. Comunmente, unha serie de tempo (serie temporal) defínese como o conxunto de observacións dunha variable aleatoria medida secuencialmente ao longo do tempo, é dicir, como a traxectoria dun SP. En xeral, a mostraxe pode ser igualmente espazada, mediante intervalos de tempo constantes, ou pode non establecerse un patrón regular neste labor de adquisición.

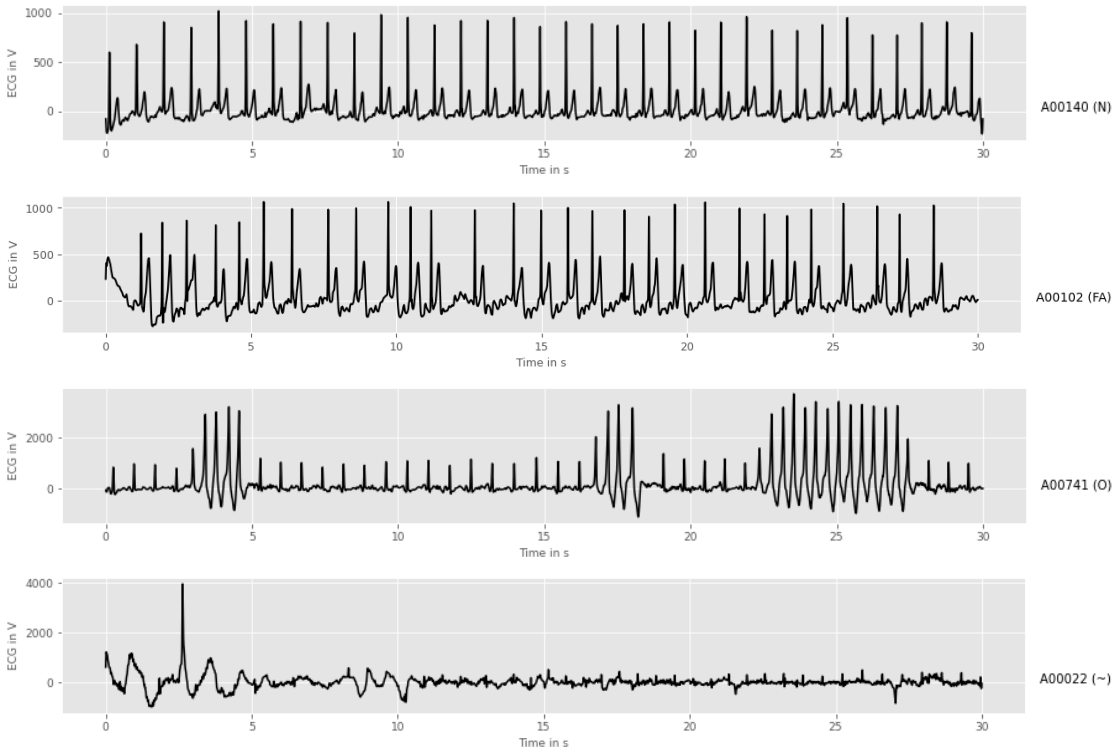


Figura 1.2: Rexistros de exemplo das catro etiquetas (**N**, **FA**, **O**, ∽).

na execución de *Construe*, marco de traballo que introduciremos na penúltima sección e que será imprescindible na tarefa de clasificación. Finalmente, cando o problema se simplifique ao cribado de FA, restrinxirémonos ao caso binario, directamente relacionado con padecer ou non esta doenza.

Para rematar coa explicación da fonte de datos, obtéñense certos valores relacionados co reparto mostral das etiquetas e que serán relevantes en discusións futuras. Concretamente, calcúlase a cantidade de rexistros de cada unha das clases, denotada por n_L, e a súa proporción mostral, denotada por p_L, onde $L \in \{\mathbf{N}, \mathbf{FA}, \mathbf{O}, \backsim\}$. A continuación, recóllense os resultados obtidos:

- Normalidade: hai n_N =5525 rexistros etiquetados como **N**, abarcando unha proporción de p_N =0.6402, o que se traduce en que o 64.02 % das etiquetas son de normalidade. Consecuentemente, o inverso é p_N^{-1} =1.5621.

- Fibrilación auricular: hai n_{FA} =733 rexistros etiquetados como **FA**, abarcando unha proporción de p_{FA} =0.0860, o que se traduce en que o 8.60 % das etiquetas son de fibrilación auricular. Consecuentemente, o inverso é p_{FA}^{-1} =11.6317, moi superior ao valor acadado para **N**. A prevalencia mostral de FA (nesta fonte de datos) é igual a 0.086.

- Outras patoloxías: hai n_O =1959 rexistros etiquetados como **O**, abarcando unha proporción de p_O =0.2298, o que se traduce en que o 22.98 % das etiquetas son doutras patoloxías. Consecuentemente, o inverso é p_O^{-1} =4.3522.

- Ruído: hai $n_\backsim$ =309 rexistros etiquetados como ∽, abarcando unha proporción de $p_\backsim$ =0.0362, o que se traduce en que o 3.62 % das etiquetas son de ruído. En consecuencia, o inverso é $p_\backsim^{-1}$ =27.5923, maior cantidade acadada, dada a baixa proporción mostral asociada a esta clase.

1.4. Descrición básica dun latexo nun ECG

O noso modelo de clasificación –do mesmo xeito que o exposto en Teijeiro et al. (2018b)– ampárase na información gardada en dous conxuntos de variables extraídos automaticamente para cada referencia. Como se mencionou con anterioridade, esta tarefa realízaa o *framewok Construe*, que obtén a información relevante a través da descrición dos rexistros de ECG nos mesmos termos que manexan os médicos cardiólogos. Ademais, o proceso de recoñecemento de patróns necesita unha detección de latexos previa, tamén implementada en *Construe*, e que vén acompañada dunha caracterización rítmica e morfolóxica que enriquece aos modelos e abre paso á extracción de características.

Neste apartado, pormenorízase na estrutura morfolóxica básica dun latexo e descríbense os seus integrantes principais. Con este propósito, mencionar que a actividade eléctrica cardíaca ocasiona que no ECG se perciban unha serie de ondas moi fáciles de recoñecer e que reflexan o comportamento eléctrico de cada latexo. Tal é como se representa na Figura 1.3, defínense:

- Ondas: curvaturas características presentes no trazo do electrocardiograma e que se repiten en cada latexo. Como advertimos na Figura 1.3, as principais ondas, por orde, son: P, Q, R, S, T e U (non observable). As ondas do ECG únense a través da liña isoeléctrica.
- Segmentos: liñas que unen dúas ondas, sen incluír ningunha delas.
- Intervalos: tramos de sinal entre dúas ondas, ambas incluídas.
- Complexos: conxuntos de ondas.

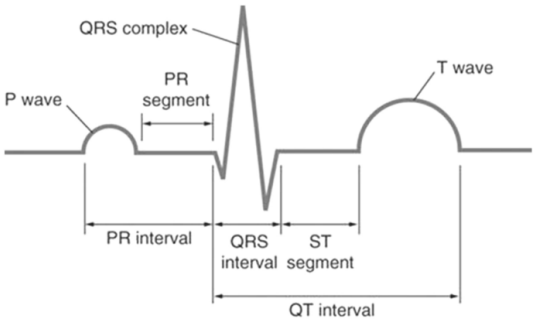


Figura 1.3: Debuxo dun latexo nun ECG con etiquetas de ondas, intervalos e segmentos, incluíndo a onda P, o complexo QRS e a onda T. Imaxe extraída do artigo de Dalal and Vishwakarma (2021).

Concretamente, defínese o complexo QRS e correspóndese coa despolarización que antecede á contracción dos ventrículos, formando unha estrutura de pico propia dun latexo de ECG. Este complexo está constituído por tres ondas, que se nomean como segue: se a primeira onda do complexo QRS é negativa, chámase onda Q, e a primeira positiva, chámase onda R (é a onda de maior tamaño); se a continuación hai outra negativa, é a onda S. Sen embargo, estas ondas non sempre son visibles e, no caso de que algunha delas non se observe, a morfoloxía do complexo QRS vese alterada[4]. De feito, se a primeira onda do complexo QRS é positiva, será a onda R e, polo tanto, non se apreciará a onda Q.

A onda P é a primeira onda do ciclo cardíaco, representa a despolarización das aurículas e está ausente na fibrilación auricular. A onda T é posterior ao complexo QRS, de menor amplitude que este, e representa a repolarización dos ventrículos. O intervalo PR é a porción do ECG que vai dende o comezo da onda P ao final da onda R. O intervalo RR é o tramo de sinal entre ondas R procedentes de

[4]No Apéndice B profundízase na explicación das morfoloxías do complexo QRS, con especial interese en aquelas detectadas por *Construe*, e introdúcese a terminoloxía usada para denotar o tamaño e ausencia das ondas que o conforman.

dous latexos consecutivos, está caracterizado pola distancia entre as ondas R e a súa duración depende da frecuencia cardíaca. Analogamente, defínense o resto de intervalos representados na Figura 1.3, que inclúe o nome das ondas, intervalos e segmentos máis importantes dun latexo nun ECG, para coñecer que medidas imos ter en conta na construción das variables globais e secuenciais. En relación aos segmentos, e considerando un último exemplo, o segmento PR correspóndese coa liña que une a onda P e a onda R, sen incluír ningunha delas.

1.5. *Construe*

Construe defínese como un marco de razoamento abdutivo baseado en coñecemento (*knowledge-based abductive framework*), destinado á interpretación automática de series de tempo. Caracterízase por ser un conxunto de métodos computacionais formado por diversos algoritmos que, agregados, implementan un ciclo de hipótese–test para proporcionar un modelo de construción da mellor explicación á evidencia dispoñible nun rexistro de ECG, compatible cos patróns de arritmias descritos na bibliografía clásica de electrocardiografía (emulando o coñecemento médico dun cardiólogo). A súa explicación completa recóllese no artigo de Teijeiro and Félix (2018) e o material necesario para a súa descarga –xunto cun manual de instalación e unha breve descrición do mesmo– atópase no enderezo

https://citius.usc.es/transferencia/software/construe,

albergado na páxina oficial do Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), centro de adscrición dos autores do traballo. Adicionalmente, neste repositorio tamén se inclúe un algoritmo destinado á clasificación automática de latexos en sinais de ECG, do que igualmente faremos uso. No que se refire á linguaxe de programación, todo o código do *software* está en `Python` (2022).

Inicialmente, o labor de detección de latexos realízase cunha versión dun algoritmo baseado na exploración de complexos QRS, exposto en Teijeiro et al. (2018a). O seu fundamento reside na modificación da aplicación `gqrs` (Goldberger et al., 2000), que xera un arquivo de anotacións cos latexos encontrados, e pretende mellorar a súa detección en presenza de sinal ruidosa, con multitude de artefactos mostrais e rexistros de baixa amplitude. Deste xeito, a evidencia inicial consiste nas anotacións obtidas para cada latexo coa aplicación `gqrs`, que proporciona unha delimitación da onda P, do complexo QRS e da onda T, información que a *posteriori* se refina.

A continuación, lévase a cabo o ciclo de hipótese–test característico de *Construe*, cunha elaboración dinámica de interpretacións mediante a combinación de diferentes modos de razoamento abaixo–arriba e arriba–abaixo, guiados por un mecanismo inspirado na atención humana, perfeccionando as anotacións existentes e mellorando a detección de latexos en fragmentos ruidosos. Este mecanismo, denominado foco de atención, contén a seguinte observación que deberá ser explicada, ou a seguinte predición derivada dunha hipótese que deberá ser comprobada. En base á natureza do contido do foco de atención, o esquema de razoamento aplicado varía do seguinte xeito:

1. Se o foco de atención contén unha observación sen explicar, realízase un proceso de abdución para conxecturar unha nova hipótese e, acto seguido, incorpórase ao foco de atención.

2. Se o foco de atención contén unha hipótese que explica unha ou varias observacións, lévase a cabo un proceso de dedución, obtendo unha predición a partir desa hipótese e do patrón de abstracción que a soporta. Seguidamente, incorpórase ao foco de atención.

3. Se o foco de atención contén unha predición feita a partir dalgunha hipótese, e existe unha observación consistente con ela, dita observación considérase explicada pola hipótese mediante un procedemento de subsunción. Acto seguido, recupérase o foco de atención anterior, correspondente á hipótese na que se subsumiu a observación, permitindo a obtención dunha nova predición.

4. Se o foco de atención contén unha predición para a cal non existe unha observación consistente, esa predición pasará a ser unha hipótese situada no foco de atención, polo que nos seguintes pasos buscarase a evidencia que soporte esta nova hipótese a un menor nivel de abstracción.

Chegados a este punto, débese salientar que, en paralelo á descrición morfolóxica e rítmica do comportamento da actividade eléctrica cardíaca, a abdución é de gran axuda para detectar e corrixir rexistros invertidos[5], froito de realizar unha incorrecta medición do sinal.

En última instancia, o resultado do preprocesado proporciona a información necesaria para a clasificación, dotando aos modelos de dous conxuntos de variables significativas, un deles relativo a características individuais dos latexos, de tamaño 83, e outro a medicións resumo do trazo completo, de tamaño 42. Para cada rexistro, estes datos almacénanse en dúas estruturas separadas: unha matriz que contén a información secuencial –onde as filas se asocian cos latexos individuais– e un vector que contén a medición das variables globais. A idea é usar esta información para axustar os criterios subxacentes ao conxunto de adestramento, modelando o coñecemento que manexan os manuais de electrocardiografía, para tomar unha decisión sobre a condición de normalidade/anormalidade dun rexistro. Por este motivo, entre as variables extraídas, inclúense propiedades morfolóxicas relacionadas coa estrutura dos latexos, medidas estatísticas sobre o intervalo RR e información espectral. No Apéndice B preséntase unha listaxe e descrición detallada das variables, con especial atención á caracterización rítmica e morfolóxica e ao preprocesado inmediatamente anterior á súa entrada nos modelos.

1.6. Ferramentas informáticas

Toda a parte experimental do presente traballo se realizou nunha computadora de escritorio con procesador Intel®Core™ i5-4590 de 16GB de memoria RAM, baixo o sistema operativo `Ubuntu` 20.04.1 LTS de 64 bits. Tanto para a representación gráfica como para a aprendizaxe, optimización de hiperparámetros e cálculo de métricas utilizadas como medidas de calidade da clasificación, empregamos as linguaxes de programación `Python` (2022) e `R Core Team` (2022) –ambas de alto nivel e orientadas a obxectos–, adicándose esta última só a algunha tarefa illada de representación.

Por unha banda, entre as librarías de `Python` (2022) máis destacables atópase *numpy* (abreviatura de *numerical* ***`Python`***), adicada á creación de vectores e matrices e á incorporación dunha gran colección de funcións matemáticas e operacións entre elas, e *pandas* (abreviatura de *panel data*), que ofrece estruturas de datos e operacións para manipular táboas numéricas e series temporais. Ademais, utilizáronse as seguintes librarías de *statistical learning* e *deep learning*: *scikitlearn* e *sklearn* (para modelos de clasificación e optimización de hiperparámetros), *TensorFlow* (creada por Google Brain e que en 2015 se presentou baixo unha licencia de código aberto, coa publicación do artigo de Dean and Monga (2015), para o manexo de redes neuronais) e *Keras* (adicada ás redes neuronais e con soporte en *TensorFlow*). Por último, as librarías destinadas á xeración de gráficas foron *matplotlib* e *seaborn*.

Por outra banda, a libraría de `R` máis destacable foi ***`R.matlab`***, utilizada na lectura e manipulación dos rexistros de ECG –gardados co formato contedor de datos binarios .mat propio da plataforma de programación e computación numérica `MATLAB`–. Para o manexo dos rexistros de ECG como series de tempo recorremos ás librarías *fpp2* e *tseries*; e para a visualización gráfica avanzada decantámonos polas librarías *ggplot2* e *viritis* (punteira na creación de escalas de cores).

[5]Moitos rexistros da base de datos estaban invertidos (procedendo de medicións da derivación RA-LA, e non da derivación LA-RA) porque o dispositivo de medida non requiría que o usuario o colocara nunha orientación particular.

Capítulo 2

Formulación da solución e métodos

Neste capítulo descríbense os modelos de clasificación utilizados no traballo, introducindo os mecanismos de selección de variables aplicados e explicando a natureza da elección da arquitectura do clasificador secuencial e do clasificador final, enfatizando a interpretabilidade dos resultados da clasificación. Primeiramente, estúdase a base teórica do modelo global, para logo traballar co modelo secuencial e, seguidamente, co clasificador *ensemble* e os criterios externos á aprendizaxe engadidos. Por último, coméntase como se levou a cabo o adestramento dos modelos e a optimización dos hiperparámetros, con resultados gardados en táboas no Apéndice C.

Seguindo a idea exposta en Teijeiro et al. (2018b), aplícase un método de *ensemble*, combinando un clasificador global e un secuencial, apoiándonos na construción dunha arquitectura de rede neuronal recorrente adaptada á natureza do problema, e que mellorará os resultados acadados no estudo anterior. En particular, unha das contribucións novas deste traballo reside en propoñer un novo clasificador secuencial e *ensemble*, sen alteracións na elección do modelo global, pero abreviando notoriamente o espazo de características, aplicando unha serie de procesos de selección de variables que, axeitadamente, reducen a complexidade dos modelos e o risco de padecer sobreaxuste (*overfitting*).

Nun primeiro paso, e a partir dos rexistros de ECG, o *framework Construe* realiza a tarefa de extracción das variables globais e secuenciais explicada na Sección 1.5, proporcionando a entrada dos modelos de clasificación global e secuencial, respectivamente, que a súa vez producen dúas saídas independentes que, ensamblándoas, constitúen a entrada do clasificador *ensemble*. Unha versión modificada deste último, incorporando criterios externos á aprendizaxe e que constituirá a versión final do noso modelo, é a responsable última da etiqueta final. Destacar que esta estratexia de *ensemble* modificada, representada no Diagrama 2.1 e aclarada na Sección 2.3, permitirá aproveitar as saídas dos clasificadores de nivel inferior para mellorar a súa xeralización e rendemento individual.

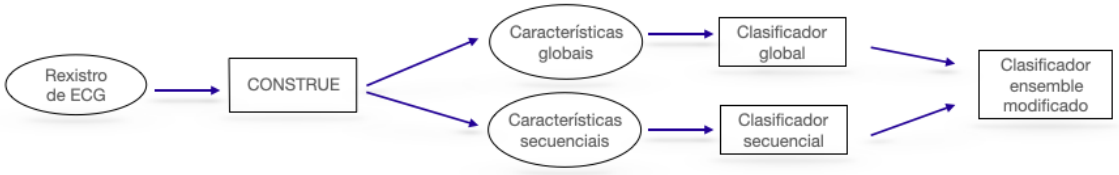


Diagrama 2.1: Arquitectura do clasificador final: métodos e datos utilizados.

Por unha parte, o clasificador global utiliza medidas resumo de todo o rexistro para decantarse por unha clase ou outra, asignando probabilidades ás mesmas e elixindo a máis verosímil (aquela cuxa probabilidade asociada sexa maior) en base ao total da secuencia observada. Isto permite detectar unha patoloxía duradeira no tempo, como o flúter ventricular, ou tal que sexa a súa persistencia o

que ocasione a anomalía, como a taquicardia e a bradicardia. Por outra parte, o clasificador secuencial emprega información de cada latexo, concentrándose na conduta temporal, descartando a normalidade por cambios de ritmo e morfoloxía, ou pola presenza de incidentes patolóxicos illados. Como consecuencia, detéctanse anomalías curtas, como os bloqueos de rama ou as extrasístoles, proporcionando unha clasificación alternativa á global que, ensamblándoa, produce unha clasificación final que utiliza as probabilidades de clases de ambas para mellorar o seu desempeño particular.

Como xa se comentou, o actual capítulo está estruturado como segue: na Sección 2.1 explícase o clasificador global e preséntase a súa base matemática, centrándonos na selección de variables e na optimización dos hiperparámetros. Na Sección 2.2 séguese un guión similar para o clasificador secuencial, coa incorporación dun novo algoritmo de selección de variables no eido da clasificación de series temporais, encamiñándonos á ilustración da auto-explicabilidade dos resultados e proporcionando varios exemplos ilustrativos da mesma. Por último, na Sección 2.3 culmina a presentación dos clasificadores integrantes do modelo proposto, coa formalización do clasificador *ensemble* e a precisión dos criterios externos á aprendizaxe engadidos, e que completan a definición do modelo final.

2.1. Clasificación global

As árbores de decisión (Quinlan, 1986) conforman modelos de aprendizaxe estatística básicos que, pese a baixa capacidade preditiva, son o sustento doutros máis refinados, tratándose dos algoritmos máis utilizados na toma de decisións na maioría de ramas da IA. O seu fundamento apóiase en estratexias de mellora incremental a partir dos datos, buscando patróns relevantes para proporcionar unha boa clasificación ou predición. Estes modelos implican a división do espazo de características en rexións simples, que determinan a súa saída: a predición dunha nova observación acostuma realizarse mediante a media (regresión) ou a moda (clasificación) das observacións de adestramento da rexión á que pertence. En xeral, se a árbore é moi grande (ten moitas ramificacións), o modelo será demasiado complexo e tenderá a sobreaxustar os datos de adestramento. Contrariamente, se é moi pequena, o modelo será máis estable pero a súa capacidade preditiva diminuirá notablemente. Intuitivamente, a profundidade das árbores será clave. O clasificador global que usaremos no noso problema apóiase no uso de moitas árbores de decisión con escasa capacidade preditiva (só un pouco superior á dun preditor aleatorio) para obter un modelo combinado cun rendemento aceptable.

Seguindo a idea propia de Teijeiro et al. (2018b), o clasificador global utilizado é o XGBoost (*extreme gradient boosting*), exposto en Chen and Guestrin (2016). Este modelo aplica a idea de impulso (*boosting*), consistente nunha aprendizaxe lenta, fundamentada na combinación de modelos con baixa capacidade preditiva (*weak learners*) para, impulsando o seu desempeño, conseguir un mellor preditor. O propósito é obter unha clasificación final como combinación ponderada das decisións individuais, segundo a exactitude das predicións, de tal xeito que se orixina un proceso iterativo de agregación de clasificadores débiles e modificación de pesos en función dos datos mal clasificados en pasos anteriores, dotándoos de máis importancia para recoller o esmero de clasificadores futuros e mellorar o desempeño final. En particular, trátase dunha mellora do método *gradient boosting* (Friedman, 2001), que contempla a opción de engadir termos de regularización, penalizando complexidade, e a poda de árbores cuxas ramificacións non favorezan á aprendizaxe.

Para presentar o modelo de maneira formal, adaptaremos o artigo orixinal de Chen and Guestrin (2016) á clasificación multiclase. En primeiro lugar, tense en conta o Algoritmo 1[1] que, aínda que está proposto para problemas de regresión, ten a clave para realizar a extrapolación desexada. A idea deste último é ensamblar árbores de decisión para, mediante unha aprendizaxe incremental, obter un modelo de regresión que proporcione un bo axuste e xeralización (o que xustifica o uso de hiperparámetros).

[1] O Algoritmo 1 está extraído do Capítulo 8 do libro de James et al. (2014) [Sección 2].

Algoritmo 1 : Construción incremental dun modelo de regresión baseado en árbores de decisión.

1. Considérase o conxunto de adestramento $\{(\overrightarrow{x_i}, y_i) \in \mathbb{R}^D \times \mathbb{R}\}_{i=1}^n$, onde $\overrightarrow{x_i} = (x_{i1}, ..., x_{iD}) \in \mathbb{R}^D$ é un vector de observacións das D variables explicativas, que constitúen un vector aleatorio $\overrightarrow{X}$, e y_i é a i-ésima observación da variable resposta Y (continua).
2. Defínese o preditor inicial, $\hat{f} \equiv 0$.
3. Fíxase o número de árbores e a súa profundidade, denotados por M e d, respectivamente.
4. Inícianse os residuos, igualándoos ás observacións: $r_i = y_i$.
5. Para cada iteración $m \in \{1, ..., M\}$:
   5.1. Constrúese unha árbore $\hat{f}^{(m)}$ de profundidade d ($d + 1$ nodos terminais), coas variables explicativas orixinais e tomando como variable resposta, os residuos da iteración anterior. É dicir, constrúese a árbore en base ao novo conxunto de adestramento $\{(\overrightarrow{x_i}, r_i) \in \mathbb{R}^D \times \mathbb{R}\}_{i=1}^n$.
   5.2. Actualízase o preditor sumando un versión reducida da nova árbore:
   $$\hat{f} = \hat{f} + \eta \hat{f}^{(m)},$$
   onde $\eta > 0$ é a taxa de aprendizaxe (*shrinkage parameter*), que reduce o impacto da resposta dunha árbore no modelo, deixando marxe de mellora para que árbores futuras poidan realizar a súa contribución individual: determina a velocidade da aprendizaxe no sentido de que regula a mellora incremental do preditor $\hat{f}$ ao engadir unha nova árbore.
   5.3. Actualízanse os residuos: $r_i = r_i - \eta \hat{f}^{(m)}(\overrightarrow{x_i})$.
6. O preditor final establécese igual a $\hat{f}$.

Recordando o modelo de regresión lineal (múltiple) clásico e a súa coñecida extensión á regresión loxística e, en xeral, aos modelos lineais xeralizados, conséguese un razoamento similar para a adaptación do Algoritmo 1 ao ámbito da clasificación multiclase. A diferenza do conxunto de adestramento definido no Paso 1, debe considerarse o determinado por

$$\{(\overrightarrow{x_i}, y_i) \in \mathbb{R}^D \times \{1, ..., K\}\}_{i=1}^n,$$

onde $\overrightarrow{x_i} = (x_{i1}, ..., x_{iD}) \in \mathbb{R}^D$ é, novamente, un vector de observacións das D variables explicativas, que constitúen un vector aleatorio $\overrightarrow{X}$, e y_i é a i-ésima observación da variable resposta Y (categórica) que, por conveniencia, establecemos que tome valores no conxunto $\{1, ..., K\}$.

A partir do conxunto de adestramento, a idea para construír unha árbore de clasificación é que cada nodo s da árbore represente unha rexión R_s do espazo de variables, que contén n_s observacións. Deste xeito, obtense unha partición do espazo preditor en J rexións disxuntas $R_1, R_2, ..., R_J$ (rectángulos), e defínese a proporción de observacións mostrais da clase k no nodo s como:

$$\hat{p}_{sk} = \frac{1}{n_s} \sum_{\overrightarrow{x_i} \in R_s} \mathbb{I}(y_i = k) \in [0, 1], \quad k \in 1, ..., K, \tag{2.1}$$

onde a función $\mathbb{I}(\cdot)$ denota á función indicadora.
Desafortunadamente, resulta computacionalmente inasequible considerar cada posible partición do espazo de características en J rexións simples (e máis aínda, no caso do *boosting*, onde construiremos unha cantidade M de árbores que logo se agregarán para proporcionar a clase final correspondente).

Como solución, vamos a adoptar unha aproximación *top-down* coñecida como *recursive binary splitting*, tal que as divisións de cada árbore se realizan dende a raíz (onde todas as observacións pertencen a unha única rexión, por estar o espazo de atributos sen dividir) aos nodos terminais, e en cada paso engádense dúas novas particións, tendo en conta a mellor división nese paso e sen valorar que sexa a óptima en relación a divisións futuras.

Usando a notación de nodos, clasificamos as observacións no nodo s segundo a súa clase maioritaria:

$$k(s) = \text{argmáx}_k \hat{p}_{sk}. \tag{2.2}$$

Para cada $k \in \{1, ..., K\}$, as cantidades $\hat{p}_{sk}$ determinan a elección da etiqueta final e, seguindo o léxico propio dos modelos de aprendizaxe estatística, reciben o nome de pesos relativos ao nodo s (para cada clase k). A expresión (2.1) resulta moi interesante de cara á interpretación dos resultados da clasificación porque facilita as proporcións de cada clase na rexión R_s, en relación á mostra de adestramento.

No problema de clasificación en K clases, a función de perda alternativa á suma residual de cadrados –propia do contexto da regresión e inutilizable con resposta categórica– que imos aplicar na construción dun clasificador combinado f (de K árbores de decisión binarias asociadas ás posibles K etiquetas) é a entropía cruzada categórica (*categorical cross entropy*). Segundo esta nova función de perda, o erro nunha rexión R_s do espazo preditor vén dado por:

$$L_s \equiv L(Y, f(\overrightarrow{X})) = -\sum_{k=1}^{K} \hat{p}_{sk} \log(\hat{p}_{sk}). \tag{2.3}$$

Destacar que a aplicación desta función de perda logarítmica penaliza ás etiquetas incorrectas tendo en conta a probabilidade de clasificación, constituíndo un criterio de erro axeitado para o cálculo do clasificador global. Deste xeito, e para a regresión loxística con K clases, o clasificador combinado resultante, $f = (f_1, ..., f_K)$, pódese formular en función da probabilidade das clases (para un nodo):

$$f_k(\overrightarrow{X}) = \log(p_k(\overrightarrow{X})) - \frac{1}{K} \sum_{j=1}^{K} \log(p_j(\overrightarrow{X})),\ k = 1, .., K,$$

de forma que ao inverter a igualdade se poden expresar as probabilidades en función de f:

$$p_k(\overrightarrow{X}) = \frac{\mathrm{e}^{f_k(\overrightarrow{X})}}{\sum_{j=1}^{K} \mathrm{e}^{f_j(\overrightarrow{X})}} \in \mathbb{R},\ k = 1, ..., K,$$

onde $p_k = p_k(\overrightarrow{X}) = \mathbb{P}(Y = 1|\overrightarrow{X})$ é a probabilidade de elixir a clase k condicionada ao vector de atributos $\overrightarrow{X}$, e calcúlase segundo a clase modal correspondente ao nodo ao que pertence $\overrightarrow{X}$ na árbore f_k que, por fixar notación, supoñamos que é o nodo s. Como resultado, obtemos que $p_k(\overrightarrow{X}) = \hat{p}_{sk}$.

Seguidamente, detallamos a base matemática do modelo de Chen and Guestrin (2016).
En primeiro lugar, e por analoxía ao Algoritmo 1, para cada paso $m \in \{1, ..., M\}$ constrúese unha árbore para cada clase, $f_k^{(m)}$ (árbore construída na iteración m para a clase k), $k \in \{1, ..., K\}$. Como resultado, obtense unha combinación de M grupos de árbores de tamaño K, tal que no $m-$ésimo grupo hai K árbores adicadas á clasificación binaria dunha clase fronte ao resto: $f^{(m)} = (f_1^{(m)}, ..., f_K^{(m)})$.
Dito doutro xeito, cada árbore destínase a clasificar os individuos en función da súa posible pertenza a unha certa clase $k \in \{1, ..., K\}$: a clasificación multiclase resólvese con K clasificadores binarios, cun enfoque un contra o resto (*one versus the rest*).

Consecuentemente, definimos a $i-$ésima predición (con variable resposta discreta) como

$$\hat{y}_i = \phi(\overrightarrow{x_i}) = \sum_{m=1}^{M} f^{(m)}(\overrightarrow{x_i}), \text{ con } f^{(m)} \in \mathcal{F}, \tag{2.4}$$

onde $\mathcal{F} = \{f(\overrightarrow{X}) = \omega_q(\overrightarrow{X});\ q : \mathbb{R}^D \longrightarrow \{1, ..., T\},\ \omega_q \in \mathbb{R}^T\}$ é o espazo de todas as posibles árbores de clasificación en K clases[2], e determínase como aquelas árbores con estrutura (topoloxía) q, pesos dos nodos terminais ω_q e número de nodos terminais T. Como resultado, a estrutura (q) dunha árbore usa regras de decisión que permiten calcular a predición final tendo en conta a puntuación dos nodos terminais (dada polo vector ω_q): a estrutura conecta cada observación co nodo terminal correspondente e, segundo a expresión (2.2), permite calcular a clase final.

O clasificador final, ϕ, vén dado pola suma ponderada[3] das predicións de cada árbore, que se obteñen como resultado de minimizar a función obxectivo

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} L(y_i, \phi(\overrightarrow{x_i})) + \sum_{m=1}^{M} \Omega(f^{(m)}), \text{ con } \Omega(\cdot) \text{ termo de regularización} \tag{2.5}$$

e L función de perda diferenciable e convexa, dada pola entropía cruzada categórica na clasificación multiclase, e que mide a diferenza entre a predición $\hat{y}_i = \phi(\overrightarrow{x_i})$ e a resposta y_i (ambas dadas en termos de probabilidades), para cada $i \in \{1, ..., n\}$. En canto á regularización, modelada agregando a función de regularización Ω, aplicada sobre cada árbore $f^{(m)}$, engádese o sumando

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel \omega_q \parallel^2$$

á función de perda, onde T é o número de nodos terminais de cada árbore $f \in \mathcal{F}$ e $\omega_q \in \mathbb{R}^T$ é o vector de pesos relativos á estrutura q. Os hiperparámetros $\gamma > 0$ e $\lambda > 0$ serán pequenos valores prefixados, que reducirán T e ω_q, conseguindo árbores máis pequenas e controlando o sobreaxuste.
Adicionalmente, na construción de cada árbore, usarase só un subconxunto aleatorio das observacións que, a súa vez, unicamente conterá un subconxunto aleatorio de variables, de tal forma que cada árbore se atopa nun escenario distinto de clasificación, forzando á xeralización do modelo final e reducindo o posible sobreaxuste. Debemos mencionar que a optimización destes hiperparámetros se explica no Apéndice C.1, indicando o enfoque seguido e os valores finalmente utilizados.

O noso propósito é minimizar $\mathcal{L}(\phi)$ con respecto de ϕ, tendo en conta que esta está restrinxida a ser unha suma de árbores (2.4). Como podemos observar, incluímos funcións como parámetros polo que é necesaria a optimización no espazo de funcións, por ser insuficiente o espazo euclidiano.

Na práctica é inviable enumerar todas as posibles estruturas (topoloxías do mellor esquema que conforma unha árbore) q e así, todas as árbores candidatas a formar parte da solución óptima. En consecuencia, óptase por un método aditivo aproximado (*steepest descent approach*) para o cálculo de $\phi = \text{argmín}_\phi \mathcal{L}(\phi)$, tal que a cada paso se incorpora a árbore que minimice unha certa función de perda, cun algoritmo que busca a mellor división do espazo de características en canto á estrutura de dita árbore. Concretamente, se $\hat{y}_i^{(m-1)}$ é a predición do i-ésimo individuo na iteración $m-1$, debemos engadir a árbore $f^{(m)}$ que minimice a perda $\mathcal{L}^{(m)}$:

$$f^{(m)} = \text{argmín}_{f^{(m)}} \mathcal{L}^{(m)}(f^{(m)}) \ :\ \mathcal{L}^{(m)}(f^{(m)}) = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(m-1)} + f^{(m)}(\overrightarrow{x_i})) + \Omega(f^{(m)}). \tag{2.6}$$

[2]En concreto, $\mathcal{F}$ representa ao espazo CART (*Classification and Regression Trees*) para a clasificación en K clases.

[3]Na expresión (2.5), a ponderación relativa á suma que define ao clasificador final refírese á penalización que se aplica a cada árbore na súa optimización, de tal xeito que $f^{(m)}$ se obtén segundo a expresión (2.6). A idea é incorporar novos clasificadores que melloren a clasificación en rexións difíciles, regulando a súa contribución ao modelo final en función das observacións mal clasificadas durante o proceso de adestramento.

Isto significa que agregaremos a árbore $f^{(m)}$ que máis beneficie ao modelo, segundo á expresión (2.6), concretando como se realiza o paso dunha iteración (partimos da iteración $m-1$, $m>1$) á seguinte. Como se expón no artigo orixinal de Chen and Guestrin (2016), na optimización de $\mathcal{L}^{(m)}$ utilízase unha expansión de Taylor de segundo orde, perfeccionando o método de descenso do gradiente que lle dá nome ao *gradient boosting*[4], requirindo tamén o cálculo da matriz hessiana da función de perda:

$$\mathcal{L}^{(m)}(f^{(m)}) \approx \sum_{i=1}^{n}\left[L(y_i,\hat{y}^{(m-1)}) + g_i^{(m-1)}f^{(m)}(\overrightarrow{x_i}) + h_i^{(m-1)}{f^{(m)}}^2(\overrightarrow{x_i}) + \Omega(f^{(m)})\right], \text{ con}$$

$$g_i^{(m-1)} = \left(\frac{\partial}{\partial\hat{y}^{(m-1)}}L(y_i,\hat{y}^{(m-1)})\right) \text{ e } h_i^{(m-1)} = \left(\frac{\partial^2}{\partial^2\hat{y}^{(m-1)}}L(y_i,\hat{y}^{(m-1)})\right),$$

onde podemos eliminar o primeiro termo do sumatorio (para $i=1,...,n$) por non depender de $f^{(m)}$:

$$\tilde{\mathcal{L}}^{(m)} = \sum_{i=1}^{n}\left[g_i^{(m-1)}f^{(m)}(\overrightarrow{x_i}) + h_i^{(m-1)}{f^{(m)}}^2(\overrightarrow{x_i}) + \Omega(f^{(m)})\right]. \tag{2.7}$$

Para cada paso $m \in \{1,...,M\}$, e co obxectivo de construír o clasificador final ϕ, a función de perda $\tilde{\mathcal{L}}^{(m)}$ optimízase na $m-$ésima iteración. Deste xeito, a función de perda total redúcese iteración a iteración, ensamblando clasificadores débiles (árbores de decisión) para dar lugar a un clasificador final con alta capacidade preditiva e de xeralización (o modelo de *boosting*). En concreto, dada unha estrutura das árbores fixa ($q^{(m)}$), lógranse calcular os pesos óptimos dos nodos terminais ($\omega_{q^{(m)}}$) minimizando directamente a expresión (2.7), o que permite obter a árbore que se debe incorporar na correspondente iteración. Ademais, esta estrutura conséguese obter con algoritmos de división exacta e aproximada, sen necesidade de enumerar o total de posibilidades: como xa comentamos anteriormente, é posible utilizar unha aproximación fundamentada na división binaria recursiva, construíndo cada árbore dende a raíz aos nodos e elixindo en cada ramificación a mellor división.

2.1.1. Selección de variables

Chegados a este punto, é o momento de estudar cales son as variables globais utilizadas na clasificación, analizando a conveniencia ou non da súa incorporación ao modelo. Como se comenta no Apéndice B, hai 42 variables globais e 83 secuenciais, resultando crucial determinar as máis importantes. Se algunha non contén información relevante –ou resulta redundante–, non se debería incluír. Recordemos que o propósito é crear un método interpretable para a clasificación automática de rexistros de ECG e para o cribado de FA, realizando melloras con respecto ao exposto en Teijeiro et al. (2018b), entre as que destaca a simplificación do espazo de características, perfeccionando a clasificación.

Para obter o modelo "óptimo", o ideal sería realizar unha busca exhaustiva, avaliando todas as posibilidades. Sen embargo, adestrar modelos de *boosting* altamente dimensionados –e máis aínda, redes neuronais– esixe unha demanda computacional elevada, polo que é recomendable utilizar un criterio por pasos. Concretamente, optamos pola eliminación progresiva (*backward process*), comezando cun modelo con todas as variables e en cada iteración eliminando unha, segundo un criterio de saída e ata que ningunha das incluídas o verifique. No lado oposto, a selección progresiva (*forward process*) non é axeitada por traballar cun problema complicado, sendo máis razoable comezar con todas as variables, co obxectivo de eliminar poucas, que con ningunha, co obxectivo de engadir moitas.

Neste caso, a selección de variables divídese en dous pasos: o Paso 1, consistente nunha técnica xa exposta na literatura, que permite utilizar eficientemente a validación cruzada (*cross validation*, CV) para eliminar variables pouco relevantes; e o Paso 2, cunha compoñente máis subxectiva, que pretende

[4]En concreto, debemos comentar que a expansión de Taylor de segundo orde é orixinal de Friedman (2001), autor que introduce por primeira vez na literatura o modelo de *gradient boosting*.

simplificar o modelo en base á información que as variables globais comparten coas súas homólogas secuenciais, dando máis poder ás redes na construción das fronteiras de decisión finais. Concretamente, e partindo das 42 variables, aplicamos unha técnica de eliminación recursiva (*recursive feature elimination*, RFE) que, utilizando un proceso interno de CV, permite excluír as posibles dependencias e colinearidades existentes, cunha idea similar á exposta en Misra and Yadav (2020). Adicionalmente, séguese un enfoque directo (*wrapper approach*) no sentido de que se involucra ao clasificador global.

Por unha banda, o obxectivo de RFE é seleccionar variables considerando recursivamente subconxuntos cada vez máis pequenos. Nun principio, o estimador adéstrase con todas as variables e a importancia de cada unha obtense a través de calquera atributo específico computable (no noso caso, será a contribución individual á medida $F1$, definida na ecuación (1)). Logo elimínanse as variables menos importantes do subconxunto actual e substitúese polo obtido ao esquecelas. Habitualmente, só se permite eliminar unha variable en cada iteración, polo que os resultados aquí presentados tamén engaden esta restrición. O razoamento repítese ata que se obtén un número desexado de variables, ata optimizar unha determinada medida de calidade ou unha función de perda. O procedemento de elección de subconxuntos óptimos realízase para cada pregadura (frecuentemente chamada *fold*) de CV, sendo o resultado final unha medida de acordo dada polo voto maioritario nos respectivos *folds*.

No noso caso, o tamaño mostral é de 8526 referencias e o custo computacional do adestramento dos modelos de *boosting* é relativamente baixo, en comparación co doutros modelos de aprendizaxe estatística (como as redes neuronais artificiais). Por esta razón, a validación aplícase con estratificación das clases e un total de 20 divisións (20-*fold* CV). Deste xeito, fracciónase a mostra en 20 subconxuntos, analizando en cada un a variable que debemos eliminar e, finalmente, retirando do modelo á elixida en máis ocasións. O resultado final determina que, por orde de eliminación, as variables excluídas son:

t1b, **n_nP**, **n_Pxcorr**, **Psmooth**, **Pdistd**, **n_aT** e **n_Txcorr**,

cuxo significado clínico se describe no Apéndice B. En particular, variables que gardan relación coa cantidade de tempo interpretado como ritmo non regular ou con medidas relativas ao segmento RR, ocupan as primeiras posicións, a diferencia de variables candidatas ao descarte, como a media da amplitude das ondas T baixo ritmo regular ou os milisegundos transcorridos ata o primeiro latexo.
Ao calcular a matriz de correlación mostral das 42 variables, obtemos as seguintes conclusións:

- A maior correlación negativa acádase entre **tSR** e **tOR**, cun valor de -0.85. Sen embargo, a súa dependencia lineal é insuficiente en relación aos resultados de eliminación de RFE.
- A maior correlación positiva acádase entre **x_xc** e **x_rrel**, cun valor de 0.93. Novamente, RFE propón manter ambas variables no modelo global.
- A menor correlación absoluta acádase entre **n_aT** e **PNN10**, cun valor de $-1.16 \cdot 10^{-4}$. Non obstante, **n_aT** elimínase do modelo como resultado da eliminación recursiva. A razón pode ser a presencia de dependencias complexas, a súa escasa conexión coa discriminación de clases, etc.

Para reproducir distancias entre clases, na Figura 2.2 realízanse estimacións non paramétricas da densidade das dúas primeiras variables eliminadas e da variable **n_aT**, en función das etiquetas **N**, **FA** e **O**. En concreto, represéntanse histogramas con punto de anclaxe fixado na orixe e anchos de banda dados pola regra de Sturges e estimadores tipo núcleo gaussianos con ventá *plug-in* (Scott, 1992). Como resultado, obtemos que o carácter discriminatorio é moi desigual: **t1b**, e algo menos **n_aT**, presentan densidades estimadas bastante solapadas, o que xustifica a súa eliminación. Sen embargo, non ocorre o mesmo con **n_nP**, pero a súa exclusión debe xustificarse en base á información aportada ao modelo, que xa debe estar modelada mediante outras variables globais.

Adicionalmente, o método utilizado ten unha interpretación gráfica moi sinxela: podemos representar unha *serie histórica* confrontando o número de variables seleccionadas e os valores da medida

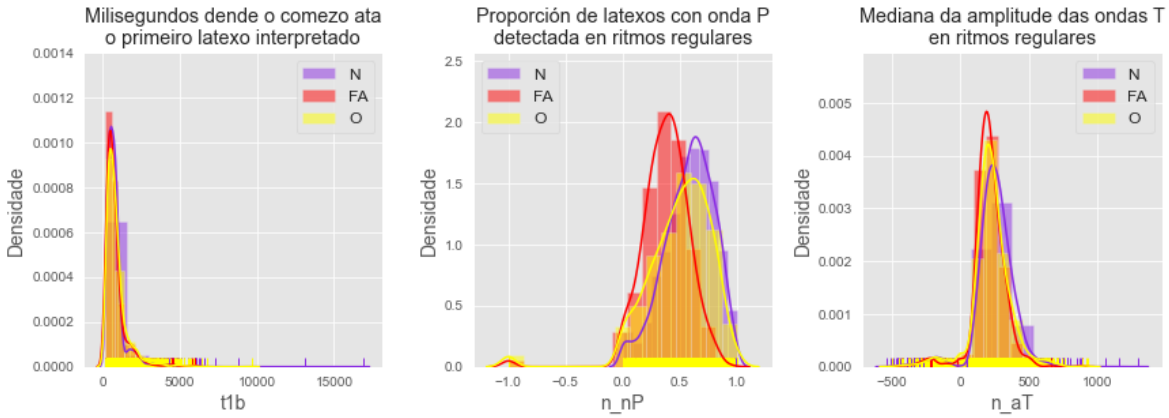


Figura 2.2: Estimacións non paramétricas da densidade das dúas primeiras variables eliminadas e de **n_aT**, segundo **N**, **FA** e **O**: histograma e estimador tipo núcleo gaussiano con ventá *plug-in* de Scott.

de calidade. Na Figura 2.3 recóllense os valores $F1$, como promedio dos resultados de 20-*fold* CV, en función do número de variables do clasificador global, eliminando aquelas propostas segundo RFE. Tendo en conta que a programación se realizou en `Python` (2022), linguaxe que comeza a enumerar os elementos dos vectores e matrices en 0, as variables globais represéntanse con índices entre 0 e 41, seguindo a ordenación presentada no Cadro B.1. Por este motivo, os valores dos eixos de abscisas desta figura varían entre 0 e 41 e 30 e 41, respectivamente, representando 41 ao modelo con todas as variables e 0 ao modelo coa variable que, individualmente, aporta máis á clasificación global.
Destacar que esta última variable se corresponde con **o_PNN50** e se explica no Apéndice B.

Na esquerda da Figura 2.3 visualizamos a representación completa e na dereita, unha ampliación, amosando os resultados das primeiras 11 eliminacións. Observamos que –con escasa diferencia do óptimo local detectado con 32 variables– o máximo global estrito se acada con 35, finalizando o Paso 1.

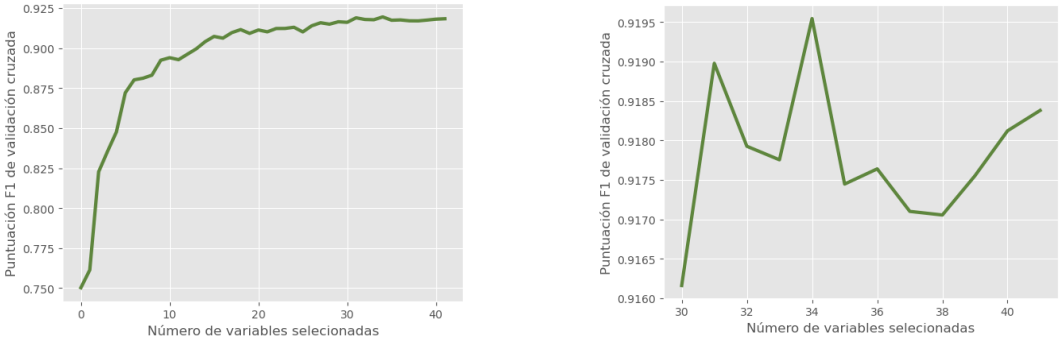


Figura 2.3: Valores $F1$ de CV segundo o número de variables do XGBoost, eliminadas en base á RFE.

En función da información modelada, no Paso 2 reduciuse máis o número de variables globais, quedándonos con 21 das 35 anteriores, eliminando aquelas que representan fenómenos pouco beneficiosos para a caracterización global, en comparación coa temporal: excluíronse variables relativas a tempos e proporcións, máximos e mínimos, medianas e variabilidade da frecuencia cardíaca. Os criterios a seguir baséanse na natureza médica das variables. No Apéndice B recóllense os resultados finais da selección de variables e a causa de exclusión das que foron descartadas, así como unha descrición das mesmas, co fin de entender que información aportan ao modelo global no caso de ser incorporadas.

Na Figura 2.4 represéntase a importancia das variables seleccionadas para adestrar ao modelo global, observando como de homoxéneo é o seu reparto no conxunto das 21 finais, destacando as relacionadas con medidas función da lonxitude do segmento RR, como **RRd_std**, **RR_Irr** e **RR**. En cada iteración do Paso 1, poderíamos representar diagramas de barras similares, observando graficamente que variables deben ser eliminadas, en función da súa contribución individual á medida $F1$ de validación promedio dos 20 *folds*. Non obstante, estas gráficas omítense por falta de espazo e porque a súa interpretación sería análoga á do diagrama de barras final aquí exposto.

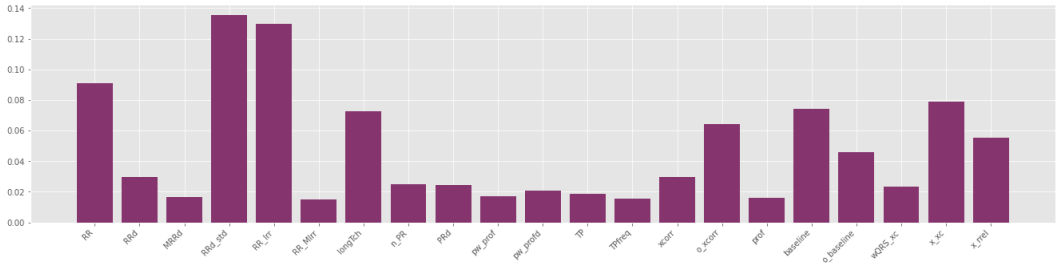


Figura 2.4: Importancia individual das variables no clasificador global final.

2.2. Clasificación secuencial

Na clasificación de secuenciais usaremos unha variante avanzada das redes neuronais artificiais (*artificial neural networks*, ANN), os mecanismos de auto-atención (*self-atention mechanisms*), que se apoian no uso de redes neuronais Bi-LSTM, isto é, redes neuronais con celdas LSTM (*long short-term memory*) bidireccionais. En concreto, conseguiremos que este tipo de redes procese o rexistro latexo a latexo, modelando dependencias a curto prazo –herdanza das redes neuronais recorrentes (*recurrent neural networks*, RNN)– e a longo prazo (resultado de engadir celdas de memoria), utilizando información pasada, presente e futura das series temporais, como son os rexistros de ECG. Por este motivo, comezamos introducindo os elementos básicos das redes neuronais, para logo presentar a súa estrutura xeral e as sucesivas modificacións que dan lugar ás redes de atención usadas no modelo secuencial.

En relación coa estrutura da rede, debemos salientar que escoller unha topoloxía concreta determina fundamentalmente o seu desempeño e resulta un problema complexo que, en certa medida, se debe intuír (en base a estruturas xa coñecidas e á natureza dos datos). Neste caso, e co fin de presentar a arquitectura da rede neuronal que conforma o clasificador secuencial, comentar que a idea dos mecanismos de auto-atención radica no uso dunha capa Bi-LSTM como antecedente a unha estrutura de atención, que proporciona unha matriz de características asociadas a cada secuencia, identificando as partes máis relevantes para a clasificación.

Diagrama 2.5: Arquitectura da rede neuronal que conforma o clasificador secuencial, baseándonos na concatenación de dúas capas de atención, co uso de redes con esquema Bi-LSTM.

Neste caso, e como se ilustra no Diagrama 2.5, a arquitectura de rede proposta contén dous camiños de atención, o primeiro destinado á análise de ritmo, e que chamaremos capa de ritmos, e o segundo

á das restantes variables extraídas por *Construe*, e que chamaremos capa xeral. O obxectivo é aplicar os mecanismos de atención para obter aqueles ritmos máis relevantes na clasificación, razón que xustifica a decisión de considerar unha capa de ritmos, e complementar a información coa das restantes características dun latexo que, segundo a súa natureza, terán máis ou menos peso na elección final.

2.2.1. Redes neuronais artificiais

Pola relevancia na historia das redes neuronais, e como motivación da súa natureza básica, comecemos coa explicación do algoritmo do perceptrón (Rosenblatt, 1961), que intenta atopar un hiperplano separador minimizando a distancia das observacións mal clasificadas á fronteira de decisión. Matematicamente, trátase dun modelo de clasificación lineal binaria no que o vector aleatorio de variables explicativas, denotado por $\overrightarrow{X} = (X_1, ..., X_D) \in \mathbb{R}^D$, se transforma mediante unha aplicación non lineal $\phi = (\phi_1, ..., \phi_M) : \overrightarrow{X} \in \mathbb{R}^D \longrightarrow \phi(\overrightarrow{X}) \in \mathbb{R}^M$ destinada a construír unha predición da forma

$$\hat{y}(\overrightarrow{X}, \omega) = f(\omega^t \phi(\overrightarrow{X})) = f\left(\sum_{j=1}^{M} \omega_j \phi_j(\overrightarrow{X})\right), \tag{2.8}$$

onde f é unha función escalonada (función de activación non diferenciable) definida como

$$f(x) = \begin{cases} 1, & x \geqslant 0 \\ -1, & x < 0 \end{cases},$$

e $\omega = (\omega_1, ..., \omega_M) \in \mathbb{R}^M$ é o vector de parámetros. Nesta situación, a predición da variable resposta, Y, denotada mediante a grafía $\hat{y}$, toma valores no conxunto $\{-1,\ +1\}$. Normalmente, o vector $\phi(\overrightarrow{X})$ contén unha compoñente $\phi_0(\overrightarrow{X}) = 1$ para modelar o nesgo, engadíndose ao sumatorio (2.8) –acompañada dun parámetro ω_0– sen máis que comezar a conta en $j = 0$:

$$\hat{y}(\overrightarrow{X}, \omega) = f(\omega^t \phi(\overrightarrow{X})) = f\left(\sum_{j=0}^{M} \omega_j \phi_j(\overrightarrow{X})\right). \tag{2.9}$$

Na práctica, disponse dun conxunto de adestramento $\{(\overrightarrow{x_i}, y_i) \in \mathbb{R}^D \times \{-1,\ +1\}\}_{i=1}^n$, onde se verifica que $\overrightarrow{x_i} = (x_{i1}, ..., x_{iD}) \in \mathbb{R}^D$ é un vector de observacións das D variables explicativas, que constitúen un vector aleatorio $\overrightarrow{X}$, e $y_i \in \{-1,\ +1\}$ é a i-ésima observación da variable resposta Y. Neste caso, e considerando que engadimos o termo escalar de nesgo ω_0, o vector de pesos $\omega = (\omega_0, \omega_1, ..., \omega_M)$ calcúlase minimizando a función de erro L_{Percep}, dada por:

$$L_{\text{Percep}}(\omega) = -\sum_{m \in \mathcal{M}} \omega^t \phi(\overrightarrow{x_m}) y_m, \text{ sendo } \mathcal{M} = \{\text{observacións mal clasificadas}\}, \tag{2.10}$$

coñecida como *perceptron loss*. O criterio de clasificación vén definido como

$$\omega^t \phi(\overrightarrow{x_i}) > 0, \text{ se } y_i = 1 \text{ e } \omega^t \phi(\overrightarrow{x_i}) < 0, \text{ se } y_i = -1, \text{ para todo } i \in \{1, ..., n\}.$$

É dicir, a clasificación binaria verifica que

$$\omega^t \phi(\overrightarrow{x_i}) y_i > 0, \text{ para todo } i \in \{1, ..., n\},$$

o que motiva o uso da función de erro dada pola expresión (2.10) como criterio de minimización. Comentar que a optimización desta función lineal escalonada en ω se pode realizar aplicando o método do descenso do gradiente estocástico[5], como se detalla no Capítulo 4 de Bishop (2006) [Sección 1.7].

[5] O nome do algoritmo inclúe o termo 'estocástico' porque se manexan aproximacións probabilísticas dos gradientes da función obxectivo con respecto ás variables de entrada, como consecuencia do ruído presente na mostra de adestramento.

Neste punto, debemos destacar que o clasificador solución do algoritmo do perceptrón presenta importantes dificultades na súa xeralización a $K > 2$ clases. Ademais baséase en combinacións lineais dunha base de funcións fixas (como se observa na expresión (2.9)), polo que pode dar lugar a malos resultados en ausencia de linealidade do espazo transformado, o que motiva a presentación de estruturas subxacentes máis complexas: as redes neuronais artificiais. O fundamento destes últimos clasificadores radica en adaptar as bases de funcións aos datos, conformando modelos hiperparametrizados que, en detrimento da interpretabilidade e do risco de sobreaxustar, proporcionan solucións moi flexibles.

O elemento de partida dunha rede neuronal artificial é o perceptrón ou neurona artificial, unidade de cálculo consistente en aplicar unha función non lineal (función de activación) a unha suma ponderada das entradas, xeralizando a idea de perceptrón que se acaba de comentar, onde f era unha función constante escalonada. O seu funcionamento, tal é como acabamos de comentar, esquematízase na esquerda da Figura 2.6. Como se expón no Capítulo 5 de Bishop (2006), as redes neuronais constrúense interconectando neuronas artificiais, agrupadas en capas. A idea é estender o modelo dado pola ecuación (2.9), impoñendo que as función ϕ_j sexan paramétricas e axustando os parámetros no adestramento, á par que os pesos $(\omega_1, ..., \omega_M)$ e o nesgo ω_0 (que agora dependerán das capas).

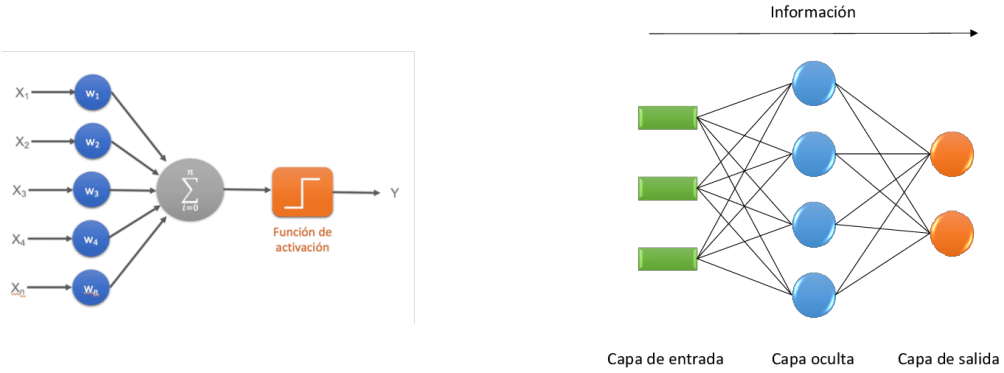


Figura 2.6: Esquerda: neurona artificial. Dereita: rede neuronal artificial (ANN).

Para a clasificación multiclase no contexto das ANN, definimos a variable resposta categórica en función da probabilidade das distintas clases. En concreto,

$$Y \in \{1, ..., K\} \text{ modélase segundo os valores que acaden os elementos } (y_1, ..., y_K),$$

onde y_k é a probabilidade asociada á clase $k \in \{1, ..., K\}$. Así pois, se a clase teórica é a $k-$ésima:

$$y_k = 1 \text{ e } y_j = 0, \ \forall j \in \{1, ..., K\} \backslash \{k\}.$$

Destacar que, necesariamente, $y_1 + ... + y_K = 1$.
Nesta situación, búscase construír unha arquitectura de rede neuronal que permita calcular predicións das probabilidades anteriores, a partir dunha mostra de adestramento, definidas como segue:

$$\hat{y}_k \equiv \hat{y}_k(\overrightarrow{X}, \omega) = \mathbb{P}(y_k = 1 | \overrightarrow{X}).$$

Con este propósito, preséntase a formalización matemática básica das redes neuronais artificiais.

Primeiramente, constrúense M combinacións lineais das variables de entrada[6] $\overrightarrow{X} = (X_1, ..., X_D)$,

$$a_j = \sum_{i=1}^{D} \omega_{ji}^{(1)} X_i + \omega_{j0}^{(1)}, \ j = 1, ..., M,$$

[6]Na notación referida ás redes neuronais artificiais, o subíndice acompaña a unha variable de entrada, e non a unha observación D-dimensional, o que xustifica o uso de letras maiúsculas.

chamadas activacións, e onde o superíndice (1) denota que se corresponden á primeira capa que –como observamos na dereita da Figura 2.6– é a capa de entrada. A continuación, calcúlanse as unidades ocultas, z_j, como a aplicación dunha función de activación diferenciable non lineal h tal que

$$z_j = h(a_j),\ j = 1, ..., M,$$

e que se corresponden coas saídas das funcións básicas da expresión (2.9). Novamente,

$$a_k = \sum_{j=1}^{M} \omega_{kj}^{(2)} z_j + \omega_{k0}^{(2)},\ k = 1, ..., K \text{ (clases)}$$

e establécese $\hat{y}_k = f(a_k)$, onde f é a función *softmax* para $K > 2$ clases[7], definida na expresión (C.1).

Do mesmo xeito que ocorría coa función de erro asociada á construción dunha árbore de decisión –que recordemos, tamén se presentaba para a clasificación en K clases na expresión (2.3), segundo a rexión á que pertence unha observación– a función de erro do modelo de rede neuronal formulado vén dada, novamente, pola entropía cruzada categórica:

$$L_{ANN}(\omega) = -\sum_{k=1}^{K} y_k \log(\hat{y}_k(\overrightarrow{X}, \omega)).$$

A arquitectura que acabamos de formalizar –representada na dereita da Figura 2.6– é a máis común na práctica, pero pode xeralizarse facilmente considerando máis capas ocultas e relaxando a restrición de interconexión das neuronas (*fully-connected layers*). Habitualmente, estas redes adéstranse co algoritmo de *backpropagation* (Rumelhart et al., 1985), unha adaptación do método do gradiente a este contexto, con moi bos resultados computacionais e que tamén se pode utilizar no adestramento das redes de atención coas que trataremos, así como no das RNN, as LSTM e as Bi-LSTM.

Como se expón formalmente no Capítulo 5 de Bishop (2006) [Sección 5.3], o algoritmo de *backpropagation* baséase na retropropagación do erro de adestramento a través da rede, dende a capa de saída á de entrada, axustando os pesos iterativamente para reducir esta medida. A clave da súa eficiencia reside na reciclaxe de cálculos realizada na avaliación das derivadas e na obtención do gradiente da función de perda, reducindo o número de operacións. Aínda así, a dificultade no adestramento das redes neuronais non só reside na súa complexidade computacional, senón que se deben ter en conta outros factores. Un bo exemplo disto atópase no Capítulo 11 de Hastie et al. (2017) [Sección 11.5], onde se analizan as redes neuronais dende un enfoque crítico, comentando que son modelos hiperparametrizados e discutindo a non convexidade da función obxectivo a minimizar na obtención dos pesos. No Apéndice C detállanse os labores adicados a reducir o sobreaxuste, dende a incorporación de criterios de regularización ata a determinación do número de unidades ocultas das capas.

Convén recordar que os rexistros de ECG son series temporais e a extracción de variables asociadas a súa clasificación secuencial se realiza latexo a latexo, o que permite asegurar a incapacidade das ANN clásicas na resolución do noso problema, ao non distribuírse no tempo. Por este motivo, presentamos as RNN, introducidas por primeira vez no artigo de Rumelhart et al. (1968), e cuxa arquitectura non só contén interconexións entre neuronas de distintas capas, senón que engade auto-conexións cíclicas para modelar dependencias temporais. Deste xeito, en cada instante de tempo, cada celda RNN recibe dúas entradas: a entrada correspondente á capa previa e a saída do instante anterior da mesma capa, permitindo incorporar a dimensión temporal (mediante a inclusión de bucles na topoloxía da rede). Na esquerda da Figura 2.7, represéntase unha neurona recorrente, que se diferencia dunha clásica na presenza dunha auto-conexión, formada por unha porta de entrada e unha de saída, ambas con función de activación continua. Con efectos ilustrativos, na gráfica dereita da Figura 2.7, a neurona recorrente

[7] En clasificación binaria, a activación na capa de saída é a sigmoide loxística. En regresión, é a identidade.

desenvólvese ao longo do tempo, o que ocasiona que a rede se desenvolva transversalmente en tantas capas como pasos temporais se dispoñan.

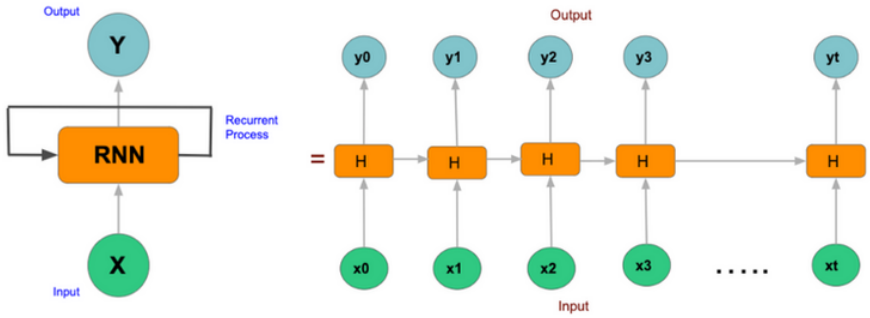


Figura 2.7: Funcionamento dunha celda RNN básica e dunha rede RNN. Imaxe extraída da páxina web https://www.theaidream.com/post/introduction-to-rnn-and-lstm.

En labores relacionados co procesado de linguaxe natural, como a tradución e transcrición de texto e fala, a cadea temporal pode ser moi longa, con dependencias entre elementos distantes críticas para a aprendizaxe, ocasionando un importante aumento no número de capas da rede desenvolta. Por exemplo, ante a problemática de traducir automaticamente unha oración extensa, a dependencia temporal entre as palabras pódese perder e, polo tanto, alterarse negativamente o significado global da sentencia. Como consecuencia, aparece o problema de decaemento do gradiente (*vanishing gradient*). Esta complicación é unha dificultade propia do adestramento de redes neuronais mediante a aplicación de métodos de descenso do gradiente estocástico e a retropropagación do erro cara atrás na rede. Durante o adestramento, os pesos actualízanse en cantidades proporcionais á derivada parcial da función de perda con respecto aos seus valores actuais. O problema resúmese na converxencia sucesiva do gradiente a valores moi pequenos, causada pola repetida aplicación da regra da cadea a funcións cuxa derivada está limitada no intervalo $[0, 1]$ (como ocorre coa función de activación sigmoide), o que impide a actualización eficaz dos pesos e, consecuentemente, o correcto adestramento da rede.

Na nosa fonte de datos, o máximo número de latexos detectados foron 173 no rexistro A01467, de 60 segundos de duración, con todos eles etiquetados como taquicárdicos: a frecuencia cardíaca é de 173 latexos por minuto, superior ao limiar de 100, a partir do cal se diagnostica taquicardia. Cunha cantidade tan elevada de latexos, a información temporal non subsiste ao longo da cadea, perdendo referencias clave na clasificación e, habitualmente, producindo solucións non desexadas. Ademais, en moitos casos, chega con secuencias de 10–12 pasos temporais para que este problema teña lugar. No artigo de Hochreiter (1998) abórdase a formulación matemática do decaemento do gradiente durante a aprendizaxe das RNN e preséntanse posibles solucións. Entre elas, atópase aplicar métodos de optimización que non usan os vectores gradiente, engadir información de orde superior (como a matriz hessiana), etc. Sen embargo, a causa da eficiencia e popularidade do algoritmo de *backpropagation*, o máis común é recorrer a redes con estruturas máis sofisticadas.

Seguindo esta última idea, e para solventar a perda de memoria a longo prazo, introdúcense as redes LSTM (Hochreiter and Schmidhuber, 1997), que permiten propagar o erro no tempo, conectando eventos correlados distanciados temporalmente. Na Figura 2.8 represéntase o funcionamento e a estrutura básica dunha celda LSTM, composta de 3 portas (1 máis que as celdas RNN) e unha (sub)celda:

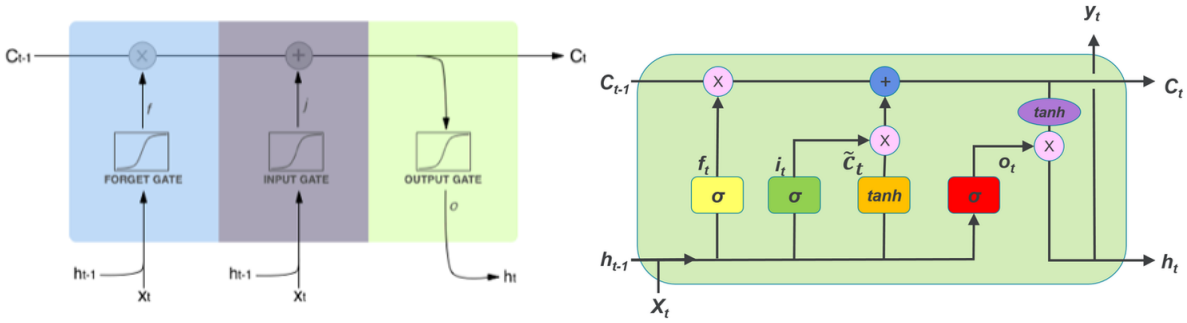
1. Unha porta de esquecemento (*forget gate*), que decide que información pasada é irrelevante nese paso temporal e, polo tanto, se debe esquecer no futuro.

2. Unha porta de entrada (*input gate*), que procesa información do estado previo e actual, onde a

información do estado previo se entende que é a procedente de pasos temporais anteriores e a actual é a información de entrada a ese paso temporal.

3. Unha porta de saída (*output gate*), que proporciona a saída final dese paso temporal, combinando o estado previo e a información de entrada e memoria actual.

4. Unha (sub)celda de memoria ou estado de celda (*memory cell or cell state*), que realiza a tarefa de esquecer o que debe ser esquecido e incorporar nova información, en función das 3 portas.



(a) Esquema das portas e subcelda.

(b) Estrutura básica.

Figura 2.8: Funcionamento dunha celda LSTM: esquema de portas e subcelda e estrutura básica. Imaxes extraídas da páxina web https://medium.com/@divyanshu132/lstm-and-its-equations.

A continuación, vamos a presentar formalmente o funcionamento básico dunha celda LSTM, co fin de concretar como *ler*, *escribir* e *borrar* información da súa memoria, conseguindo gardar dependencias temporais distantes. En primeiro lugar, debemos comentar que o instante temporal actual o denotaremos por t. A clave do funcionamento dunha celda LSTM é a (sub)celda de memoria, denotada por C_t, e que dotada de toda a información previa, permite que aquela relevante transite ao longo do tempo. Precisamente, trátase da liña horizontal que atravesa a parte superior dos diagramas da Figura 2.8 (a) e (b), dende C_{t-1} ata C_t. Por outra banda, as portas de esquecemento, de entrada e de saída –todas con función de activación sigmoide– son as encargadas de regular o fluxo de información segundo a súa transcendencia, asignando valores no intervalo $[0, 1]$.
A maiores, sexa σ a grafía que representa á función de activación sigmoide, W_x a matriz de pesos relativos á porta/(sub)celda neuronal x e b_x o nesgo relativo á porta/(sub)celda neuronal x.

O cálculo da saída da porta de esquecemento, denotada por f_t, compútase como

$$f_t = \sigma(W_f \cdot [\overrightarrow{h}_{t-1}, X_t] + b_f), \tag{2.11}$$

onde $\overrightarrow{h}_{t-1}$ é o estado anterior da celda, X_t é o vector de entrada actual[8] e W_f e b_f son, respectivamente, a matriz de pesos e o nesgo relativo á saída. Ademais, a pesar de que non aparece directamente na expresión (2.11), defínese $\overrightarrow{h}_t$ como o estado da celda LSTM actual (en tempo t). O resultado da porta de saída é unha cantidade entre 0 e 1, que modela como debemos omitir/manter a información actual e do estado previo na (sub)celda de memoria (isto é, a súa permanencia):

- O valor 0 significa que a porta bloquea o paso de información (suprímese toda).
- O valor 1 significa que a porta permite que toda a información continúe (mantense toda).

[8] Omítese incluír $\rightarrow$ ao vector de entrada en tempo t, X_t, para evitar confusións relacionadas coa notación propia das redes LSTM, onde o símbolo $\rightarrow$ se reserva para o estado da celda en tempo t, $\overrightarrow{h}_t$.

A porta de entrada, denotada por i_t, decide que información nova incorporar ao estado da celda. No seu cálculo, realízanse unha serie de operacións que se poden organizar en 3 pasos:

i. Filtrado de información pasada e actual (similar ao cálculo de f_t), regulando que información engadir á (sub)celda de memoria:

$$i_t = \sigma(W_i \cdot [\overrightarrow{h}_{t-1}, X_t] + b_i),$$

onde W_i e b_i son, respectivamente, a matriz de pesos e o nesgo relativo a este cómputo.

ii. Creación dun vector contendo todos os posibles valores a agregar ao estado da celda:

$$\tilde{C}_t = \tanh(W_C \cdot [\overrightarrow{h}_{t-1}, X_t] + b_C),$$

onde, novamente, aparecen a matriz de pesos e o nesgo, W_C e b_C. Neste caso, utilízase a función de activación tanxente hiperbólica por presentar a vantaxe de manexar con facilidade números negativos, reducindo a recta real ao intervalo $(-1, 1)$.

iii. Actualización da información da (sub)celda de memoria en base aos pasos anteriores:

$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t,$$

onde o operador $\circ$ denota á multiplicación elemento a elemento[9]. Finalmente, a porta de saída, denotada por o_t, decide que información actual utilizar como saída da (sub)celda de memoria nese instante. O estado final da celda LSTM compútase multiplicando o resultado da función de activación tanh aplicada ao estado da (sub)celda de memoria actual e á aplicación dunha activación sigmoide a información pasada e actual, formalizándose como:

$$\overrightarrow{h}_t := \overrightarrow{LSTM}(X_t, \overrightarrow{h}_{t-1}) = o_t \circ \tanh(C_t), \tag{2.12}$$

onde o_t é o resultado da porta de saída e vén dado por

$$o_t = \sigma(W_o \cdot [\overrightarrow{h}_{t-1}, X_t] + b_o).$$

Co fin de garantir a validez das expresións anteriores para calquera instante temporal $t \in \{1, ..., T\}$, débense engadir dúas restricións asociados aos valores en tempo cero:

$$h_0 = 0 \quad \text{e} \quad C_0 = 0.$$

Isto interprétase como que o estado inicial dunha celda LSTM é nulo, posto que non temos coñecemento de instantes anteriores, e o mesmo ocorre coa información da (sub)celda de memoria.

As redes LSTM constrúense como unión de celdas LSTM, igual que as redes RNN se constrúen como unión de celdas RNN e as redes ANN se constrúen como unión de neuronas artificiais básicas. Polo que acabamos de explicar, esta nova estrutura de rede neuronal soluciona a perda de memoria a longo prazo, co inconveniente de engadir complexidade aos cálculos e á interpretabilidade.

Acabamos de presentar as redes LSTM como un modelo de aprendizaxe de secuencias que incorpora información pasada e actual, con dependencias a longo e curto prazo. Sen embargo, estas redes non poden integrar eventos futuros, sendo necesario utilizar redes LSTM bidireccionais (Bi-LSTM).

[9]Dadas dúas matrices $A, B \in \mathcal{M}_{m\times n}$, $A \circ B$ é unha matriz $M_{m\times n}$ cuxos elementos veñen definidos como

$$(A \circ B)_{ij} = (A)_{ij}(B)_{ij},\ i \in \{1, ..., m\},\ j \in \{1, ..., n\}.$$

Seguindo o razoamento exposto en Schuster and Paliwal (1997) para as redes RNN, a idea é usar dúas redes LSTM, unha para propagar a información cara a diante (no sentido usual de fluxo) e outra cara atrás (no sentido oposto ao usual), e conectar ambas na capa de saída. Na Figura 2.9 represéntase o seu funcionamento básico, como unión de dúas redes LSTM. No presente traballo usaremos redes Bi-LSTM para modelar conxuntamente referencias pasadas e futuras, aproveitando toda a información contextual do rexistro de entrada.

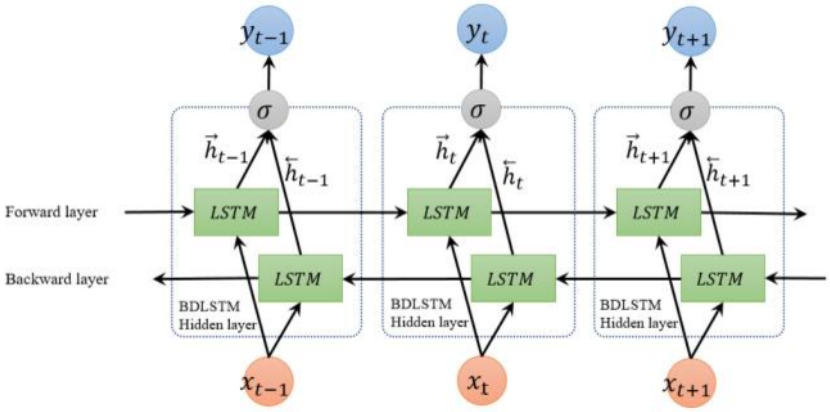


Figura 2.9: Funcionamento dunha rede Bi-LSTM. Imaxe extraída do artigo de Cui et al. (2016).

Co obxectivo que acabamos de mencionar, poderíamos modificar o clasificador secuencial exposto en Teijeiro et al. (2018b) implementando unha arquitectura de rede Bi-LSTM, como substituta da rede LSTM considerada: deste xeito, modelaríamos conxuntamente dependencias pasadas e futuras, a longo e curto prazo. Sen embargo, o resultado da clasificación sería opaco no sentido de que a rede só proporcionaría as probabilidades finais das clases, sen información adicional sobre a importancia dos pasos temporais (dos diferentes latexos) na clasificación, tratándose dun claro exemplo de modelo de *deep learning*. Co propósito de solventar esta opacidade dos resultados, e seguindo a idea exposta en Lin et al. (2017), introdúcense os mecanismos de atención na clasificación de secuencias.

Orixinalmente, as redes de atención deseñáronse para o problema de tradución de texto e fala entre idiomas, co obxectivo de aliñar e coñecer a implicación textual das diferentes palabras dunha oración no resultado final (Bahdanau et al., 2015). No noso caso, extrapolando o elaborado para a análise de texto ao contexto do *Physionet*/CinC *Challenge* 2017, os latexos serán as palabras e as oracións serán os rexistros de ECG, valorando a agregación de latexos como substituta da de palabras. A clave radica en aplicar os avances na interpretabilidade das redes LSTM e Bi-LSTM expostos en Lin et al. (2017) para crear un modelo secuencial que asigne unha cuantificación da importancia dos distintos latexos na clasificación. Neste sentido, a situación é máis semellante á análise de sentimentos en fragmentos de texto –que a partir da información recollida nun conxunto de oracións permite clasificar un texto en función do sentimento que transmite (dentro dun conxunto de emocións posibles)– que á tradución e transcrición de texto e fala, onde tanto a saída como a entrada son secuencias temporais.

A idea é usar unha capa Bi-LSTM como antecedente dunha estrutura de atención, que proporciona unha matriz de características asociadas a cada secuencia temporal (no noso caso, a cada rexistro de ECG), identificando os fragmentos máis importantes para a clasificación (cuxo peso na elección da etiqueta final é superior). Mantendo a notación exposta na explicación das redes LSTM, considérase unha secuencia de entrada con T pasos temporais

$$S = (X_1, ..., X_T)^t \in \mathcal{M}_{T \times D},$$

onde $X_t = (X_{t1}, ..., X_{tD}) \in \mathbb{R}^D$ é un vector D-dimensional coas variables de entrada do t-ésimo paso

temporal, para $t \in \{1, ..., T\}$. A partir de aí, calcúlase a saída da rede Bi-LSTM como concatenación dos estados ocultos dunha rede LSTM no sentido usual e unha no sentido oposto, ambas con u neuronas:

$$H = (h_1, ..., h_T) \in \mathcal{M}_{T\times 2u}, \text{ onde } h_t = (\overrightarrow{h}_t, \overleftarrow{h}_t),\ t = 1, ..., T,$$
$$\overrightarrow{h}_t := \overrightarrow{LSTM}(X_t, \overrightarrow{h_{t-1}}) \text{ e } \overleftarrow{h}_t := \overleftarrow{LSTM}(X_t, \overleftarrow{h_{t+1}}).$$

A continuación, co fin de codificar a secuencia S de lonxitude variable T nunha matriz/vector de dimensión fixa[10], considéranse todos os estados ocultos almacenados en H como entrada do mecanismo de atención, obtendo un vector de anotacións a. No cálculo de a úsase a función de activación *softmax* para garantir que sexa un vector unitario e, consecuentemente, obter un reparto comparable entre os distintos rexistros. Concretamente, necesitamos unha matriz $W \in \mathcal{M}_{d_a\times 2u}$ e un vector de parámetros $w \in \mathbb{R}^{d_a}$, onde d_a é un hiperparámetro que podemos fixar arbitrariamente, tales que

$$a = softmax\ (w \cdot \tanh(W \cdot H^t)) \in \mathbb{R}^T. \tag{2.13}$$

Logo resumimos os estados ocultos da rede Bi-LSTM –almacenados en H– segundo o peso proporcionado polo vector a, para conseguir unha representación vectorial m da secuencia de entrada. Normalmente este vector focalízase nun fragmento específico da secuencia, como un conxunto de pasos temporais representativos da etiqueta final. Os pesos resultantes da atención compútanse multiplicando os estados ocultos da rede Bi-LSTM polo vector de anotacións:

$$m = a \cdot H \in \mathbb{R}^{2u}. \tag{2.14}$$

Se en lugar de considerar un vector de anotacións, a, barallamos usar unha matriz de anotacións

$$A = softmax\ (W' \cdot \tanh(W \cdot H^t)) \in \mathcal{M}_{r\times T},\ \text{con } W' \in \mathcal{M}_{r\times d_a}, \tag{2.15}$$

como se ilustra no Diagrama 2.10, onde cada unha das súas r filas sexa un vector calculado de igual modo que a, obtemos unha matriz de pesos M dada por

$$M = A \cdot H \in \mathcal{M}_{r\times 2u}. \tag{2.16}$$

No caso de manexar matrices de anotacións, debe aplicarse algún criterio de regularización –como a norma matricial de Frobenius, denotada comunmente por $\|\cdot\|_F$– para evitar unha elevada dependencia lineal das filas de M e que, en consecuencia, sexan redundantes, representando información moi similar. Seguindo a idea exposta en Lin et al. (2017), establécese que

$$P = \| (AA^t - I) \|_F$$

é o termo de penalización que se engade á función de custo para dar lugar a unha perda penalizada que garante que A sexa o *máis ortogonal posible* no sentido de que AA^T sexa parecida á matriz identidade (que A^t sexa semellante a unha posible matriz inversa de A, de existir). Posteriormente, e para cada secuencia, os vectores de atención, almacenados nunha matriz de dimensión $r \times 2u$, serán as entradas dunha rede ANN que, con saída *softmax*, proporcionará a clase final.

Limitándonos ao problema presentado, a arquitectura de rede proposta contén dous camiños, o primeiro destinado á análise de ritmo, e que chamaremos capa de ritmos, e o segundo á das restantes variables, e que chamaremos capa xeral. Como se representa no Diagrama 2.5, a clasificación de secuencias benefíciase da capa rítmica no sentido de que a aparición e/ou persistencia dun ritmo no tempo

[10] Cando se menciona á codificación da secuencia S de lonxitude variable T nunha matriz/vector de dimensión fixa, estámonos referindo á matriz $M \in \mathcal{M}_{r\times 2u}$ e ao vector $m \in \mathbb{R}^{2u}$, definidos segundo as expresións (2.16) e (2.14), respectivamente. No primeiro caso, a dimensión da matriz depende do número de vectores de atención (r) e da cantidade de unidades ocultas da rede Bi-LSTM ($2u$), reducíndose a dependencia ao último valor para o caso do vector.

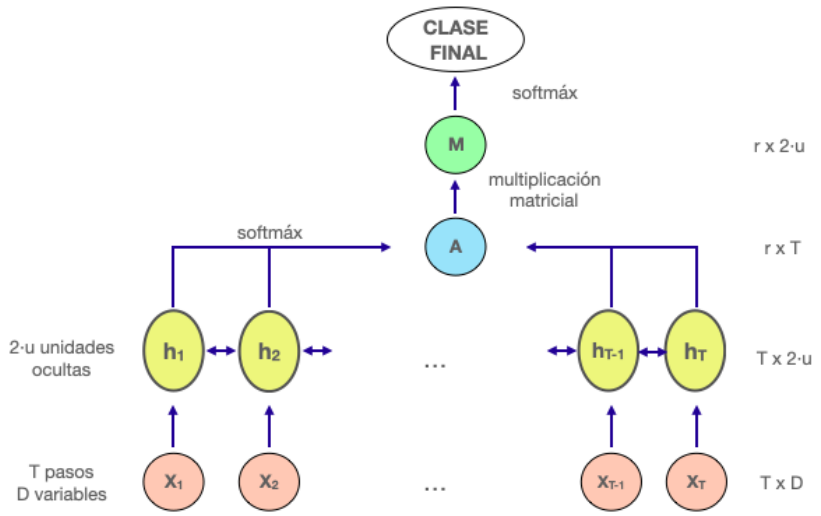


Diagrama 2.10: Arquitectura básica do modelo de auto–atención (*self–attention*).

permite afirmar e impoñer unha certa hipótese: un ritmo patolóxico condiciona a presenza dunha anomalía, así como un ritmo regular constante e non patolóxico (descartando bradicardias e taquicardias, isto é, ritmos lentos e rápidos) nos dirixe á etiqueta de normalidade. Sen embargo, esta información non é suficiente para a clasificación: o comportamento rítmico non é caracterizador dunha enfermidade, esixindo engadir unha capa coas restantes variables, incorporando información morfolóxica e espectral.

Inicialmente, e para cada camiño, unha rede Bi-LSTM proporciona a entrada ao mecanismo de *self-atention*, que calcula un conxunto de vectores de pesos operando nos estados ocultos. Este conxunto de vectores multiplícase polos estados ocultos para obter unha ponderación dos mesmos, definida como os pesos de atención. Posteriormente, e como se esquematiza no Diagrama 2.5, concaténanse para crear a entrada dunha ANN e calcúlase a clase final. No noso caso, e despois de probas de ensaio-erro, só usamos un vector de atención en cada camiño: engadir máis aumenta a complexidade e non reportaba melloras aparentes na clasificación, ademais de esixir a incorporación dun termo adicional de penalización por camiño, para garantir a mínima colinealidade posible nas matrices de atención.

En canto á regularización, extrapólase o razoamento aplicado na regresión *ridge* e *lasso* para a regresión lineal e os modelos lineais xeralizados ao ámbito das redes neuronais. Por este motivo, débense elixir os parámetros de regularización, tanto para os pesos das celdas ANN como para os pesos das celdas LSTM (para a súa saída e para a (sub)celda de memoria). Ademais, tamén é necesario determinar a taxa de aprendizaxe, o número de neuronas de cada capa da rede, etc. Comentar que a optimización destes hiperparámetros se explica no Apéndice C.2, indicando os valores finalmente empregados.

2.2.2. Selección de variables

Neste apartado motívase e formalízase o proceso de selección de variables secuenciais, preséntanse os seus resultados e proporciónase unha ordenación das mesmas en función dunha medida de relevancia final. En canto ao guión ideado, comézase coa descrición do algoritmo proposto –que é unha contribución orixinal do traballo– para posteriormente aplicalo ao problema de interese. Debemos destacar que todos os resultados acadados e que se presentarán no Capítulo 3 sobre o noso modelo son posteriores á tarefa de selección de variables e, para ambos labores, utilízase a arquitectura de rede neuronal do Diagrama 2.5. Para explicar o fundamento do algoritmo, recordemos que as redes de atención dotan á clasificación secuencial de interpretabilidade. En concreto, como resultado da súa incorporación, dis-

poñemos de vectores de pesos asociados a cada secuencia, que indican a relevancia dos distintos pasos temporais na clasificación. Por este motivo, resulta intuitivo propoñer un mecanismo de selección de variables sostido na auto-explicabilidade da solución.

En relación á formalización matemática presentada anteriormente, os vectores claves para a atención, en xeral, almacénanse na matriz de anotacións $A \in \mathcal{M}_{r\times T}$ definida na ecuación (2.15). No noso caso, almacénanse en dous vectores de anotacións $a \in R^T$ definidos segundo a ecuación (2.14) e correspondentes ás capas de atención de ritmos e xeral. Estes vectores, unitarios e de lonxitude T (compartida coa da secuencia de entrada S), modelan a importancia dos distintos pasos temporais (latexos) na elección da etiqueta final, segundo se ilustra no Diagrama 2.11.

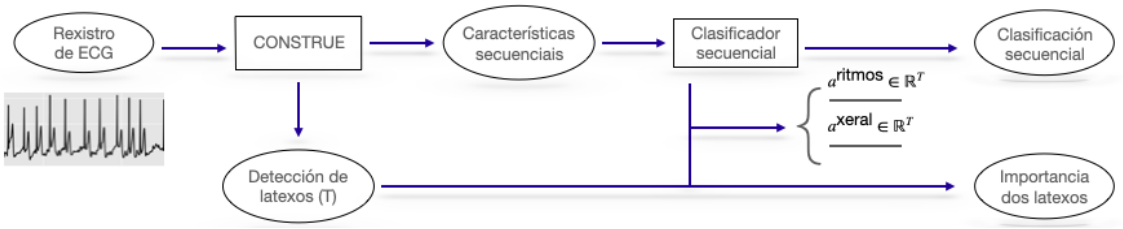


Diagrama 2.11: Importancia dos latexos no modelo secuencial en relación á arquitectura 2.5.

Por este motivo, a base do algoritmo reside na análise da influencia das variables dispoñibles no vector de atención xeral, excluíndo ritmos e morfoloxías. A idea é determinar a súa importancia a cada paso temporal, analizando como varían os pesos de atención dos rexistros ben clasificados en cada iteración, para medir a notoriedade/irrelevancia das variables. Neste aspecto, débese ter en conta a distribución das etiquetas para non eliminar variables clave na discriminación de rexistros procedentes de clases infra-representadas. Pode ocorrer que unha variable sexa fundamental para diagnosticar un episodio patolóxico propio de FA pero, se ignoramos que a porcentaxe mostral de referencias etiquetadas como **FA** é menor que a porcentaxe mostral de referencias etiquetadas como **N**, pódense obter resultados enganosos. Unha solución radica en ponderar os rexistros segundo a proporción mostral da súa clase, corrixindo nesgos prexudiciais para a discriminación de **FA** e **O**.

Algoritmo proposto

En primeiro lugar, introdúcese a notación imprescindible para a formalización da idea. Precisamente, sexa $\mathcal{T} = \{\text{características secuenciais}\}$, tal que $\mathcal{K} = |\mathcal{T}|$, e sexa $\mathcal{F} \subsetneq \mathcal{T}$ o conxunto de características fixas[11], tal que unha característica fixa é aquela que non entra no proceso de selección de variables e, previa xustificación, se engade ou desbota directamente do modelo. En consecuencia, a selección de variables aplícase ao conxunto $\mathcal{T}\backslash\mathcal{F}$, de cardinal $|\mathcal{T}\backslash\mathcal{F}| = |\mathcal{T}| - |\mathcal{F}| > 0$.

No noso caso, os ritmos, codificados como 12 variables categóricas binarias 0-1, non entran na selección de variables por tratarse, en realidade, dunha soa característica con relevancia médica en si mesma, que recordemos que define unha capa de atención específica. As 28 morfoloxías –que polo mesmo razoamento tamén se poden entender como una soa variable– exclúense directamente do modelo e nin sequera se chegan a incorporar ao conxunto $\mathcal{F}$, composto só polos 12 ritmos, logo de determinar que non aportaban información substancial[12]. En consecuencia, aplícase a selección de variables ao conxunto $\mathcal{T}\backslash\mathcal{F}$, con $|\mathcal{T}\backslash\mathcal{F}| = 55 - 12 = 43$.

[11] En xeral, o conxunto de características fixas, $\mathcal{F}$, pode ser baleiro pero, obrigatoriamente, debe ser un subconxunto estrito de $\mathcal{T}$. Se $\mathcal{F}$ e $\mathcal{T}$ coincidiran, non sería posible realizar ningún tipo de proceso de selección de variables.

[12] No Capítulo 4 discútese en máis detalle a exclusión das morfoloxías no modelo final, presentando os inconvenientes da súa incorporación e analizando diferentes modificacións da arquitectura presentada no Diagrama 2.5.

Seguidamente, e cun enfoque xeral externo á clasificación automática de rexistros de ECG, no Algoritmo 2 enuméranse e explícanse as etapas do selector proposto. Entre as súas principais propiedades, destacar que se inclúe o uso da validación cruzada de k pregaduras ($k-fold$ CV) para asegurar unha boa xeralización dos resultados, como se realiza no adestramento. Ao igual que na selección de variables globais, trátase dun enfoque *backward*, por razóns similares.

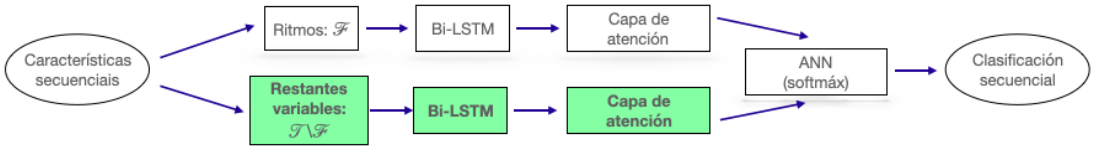
Algoritmo 2 : Selección de variables secuenciais baseándonos nos mecanismos de atención.

Para cada *fold* de validación cruzada, realízanse as seguintes etapas:

1. Adestrar á rede coas $\mathcal{K}$ variables iniciais, segundo a arquitectura do Diagrama 2.5, durante un número prefixado de iteracións, ata obter un valor razoable dunha medida de calidade da clasificación ou un valor da perda total aceptablemente baixo. O obxectivo é eliminar o nesgo ocasionado pola inicialización ao azar dos pesos e dotar de robustez ao algoritmo. Ao finalizar esta etapa, os pesos son distintos aos inicializados ao azar*.

   *É moi importante non comezar con pesos aleatorios e que estes xa tomen valores froito dun adestramento previo. Noutro caso, a aleatoriedade da asignación podería dar lugar a resultados erróneos, propios dunha rede mal adestrada. Polo xeral, os valores iniciais dos pesos elíxense próximos a 0, mediante algún procedemento aleatorio. Isto ocasiona que a aplicación da función sigmoide (función de activación de moitas das capas) teña un comportamento aproximadamente lineal, incrementándose a non linealidade da mesma a medida que os pesos aumentan.

2. Realizar CV na capa de atención e só unha iteración, para determinar que variable eliminar. Este enfoque, comparado coa CV aplicada a toda a rede, reduce considerablemente o custo computacional. A continuación, amósase un esquema da arquitectura do clasificador secuencial, destacando en cor verde a capa de atención xeral, clave na execución do algoritmo.



   Para cada $j \in \mathcal{T}\backslash\mathcal{F}$, considerar os pesos relativos á capa de atención, que definen un vector numérico a_j^1 (partindo dos pesos finais da Etapa 1) coas variables $(\mathcal{T}\backslash\{j\})\backslash\mathcal{F}$. Eliminar a variable j_1 que, entre todos os rexistros ben clasificados, provoque menos cambios na asignación de pesos. Calcular os pesos finais (necesarios en etapas futuras) como aqueles obtidos ao realizar unha iteración máis no adestramento da rede, tomando como punto de partida a Etapa 1 pero utilizando só as variables $(\mathcal{T}\backslash\{j_1\})\backslash\mathcal{F}$ nesta última iteración.

   Concretamente, debe definirse a importancia dunha variable na asignación de pesos en función da súa aportación á clasificación final dos rexistros ben clasificados. Isto é, debe identificarse aquela variable que provoque menos variación no vector de atención xeral cando non se omite do adestramento e establecer que esa é a variable candidata ao descarte.

3. Repetir a Etapa 2 ata satisfacer un criterio de parada preestablecido**. En xeral, esta etapa constará de varios pasos ou iteracións, aos que nos referimos como Paso 3.i con $i \geqslant 1$, sendo i o número de veces que se repite a Etapa 2. Entre os criterios a manexar, pódese pensar en eliminar un determinado número de variables, terminar cando os pesos discrepen máis dun certo limiar ao eliminar calquera das variables restantes, cando a perda total aumente demasiado ou cando algunha medida de calidade empeore considerablemente. Na clasificación multiclase, cunha base de datos desbalanceada, suxírese utilizar a medida $F1$.

   Defínese $\mathcal{K}^i$ como o número de variables candidatas ao descarte no Paso 3.i. No noso caso, $\mathcal{K}^1 = 41$ porque dispoñemos de 41 variables na primeira iteración, é dicir, no Paso 3.1 (como

consecuencia de que na Etapa 2 xa se eliminou unha variable). Por outra banda, a ordenación das variables en cada paso desta etapa segue un razoamento común en todos eles: comezando co Paso 3.1 (análogo para os restantes), e para cada variable $j \in (\mathcal{T}\backslash\{j_1\})\backslash\mathcal{F}$, calcúlase o vector de pesos coas variables de $(\mathcal{T}\backslash\{j_1,\ j\})\backslash\mathcal{F}$ e compárase co obtido coas variables de $(\mathcal{T}\backslash\{j_1\})\backslash\mathcal{F}$. Novamente, elimínase a variable j_2 que, entre todos os rexistros ben clasificados, provoque menos cambios na asignación de pesos.

**Na Etapa 2 adéstrase a rede dende cero. Sen embargo, os pesos iniciais non son aleatorios senón que son os proporcionados pola Etapa 1. No Paso 3.1, os pesos iniciais son os pesos finais da Etapa 2. Finalmente, para cada Paso 3.i, os pesos iniciais son os pesos finais do paso anterior.

4. Gardar a arquitectura de rede neuronal construída coas $\mathcal{K}' < \mathcal{K}$ variables restantes (conxunto final de variables), adestrar a rede dende cero e proporcionar a clasificación secuencial final.

Unha vez presentado o Algoritmo 2, cómpre realizar unha serie de comentarios importantes sobre a súa implementación práctica e rendemento:

i. Implementación. Para que a estrutura da rede sexa estable, o proceso de eliminación do algoritmo efectúase fixando os valores da variable eliminada a 0, para todos os pasos temporais. Ao rematar o proceso iterativo, créase unha nova estrutura de rede neuronal, coa mesma natureza que a orixinal, pero sen considerar as variables eliminadas. A idea é evitar incluír variables con todos os seus valores redefinidos a 0 porque só aportaría complexidade e ningunha vantaxe discriminatoria.

ii. Rendemento. Débese analizar a conduta do Algoritmo 2 e determinar se é homoxénea nos *folds* co obxectivo de sopesar a súa validez como mecanismo xeral de selección de variables, aplicable a outras bases de datos e en contextos externos á electrocardiografía.
En caso de pequenas discrepancias, pódense considerar criterios de acordo como a opinión maioritaria, acadando a conformidade de todos os *folds* para propoñer un conxunto de variables candidatas ao descarte e un conxunto final de variables utilizadas no adestramento da rede neuronal que, na notación presentada, será $\mathcal{K}'$.

A continuación, recóllese unha reflexión relativa ao cálculo da noción de importancia mencionada no Algoritmo 2, destinada a aclarar como obter medidas numéricas que nos permitan decantarnos por unha variable ou outra (en relación a súa eliminación). Con este propósito, recordemos que o tamaño mostral en problemas de aprendizaxe estatística é, habitualmente, moi elevado. Deste xeito, non resulta lóxico intentar realizar unha comparación individual dos pesos de todos os rexistros unha vez eliminada cada unha das variables (na Etapa 2 e en cada paso da Etapa 3). Para o noso caso, deberíamos comparar $n = 8626$ vectores de atención (de lonxitude variable, en función da secuencia inicial), o que resulta inviable e motiva definir unha medida de peso global.

Para cada rexistro $r \in \{1, ..., n\}$, sexa a_r (no noso caso, $a_r \equiv a_r^{\text{xeral}}$) o vector de pesos de atención obtido coas $\mathcal{T}\backslash\mathcal{F}$ variables e a_r^{-j} o vector de pesos de atención obtido coas variables $(\mathcal{T}\backslash\{j\})\backslash\mathcal{F}$.
Na Etapa 2, elimínase aquela variable $j_1 \in \mathcal{T}\backslash\mathcal{F}$ tal que

$$j_1 = \text{argmín}_{j\in\mathcal{T}\backslash\mathcal{F}} \sum_{r\in G_1} \frac{1}{p_r} ||a_r - a_r^{-j}||_2^2, \tag{2.17}$$

onde

$$G_1 = \{r \in \{1, ..., n\} :\ \text{o rexistro } r \text{ está ben clasificado na Etapa 1}\},$$

p_r é a proporción mostral da clase do rexistro r,

$$||a_r - a_r^{-j}||_2^2 = \sum_{k=1}^{l_r} (a_{rk} - a_{rk}^{-j})^2$$

é o cadrado da distancia euclidiana entre os vectores $a_r = (a_{r1}, ..., a_{rl_r})$ e $a_r^{-j} = (a_{r1}^{-j}, ..., a_{rl_r}^{-j})$, e l_r é a lonxitude da secuencia r, é dicir, o número de observacións da serie de tempo asociada. Usamos o cadrado da norma $\mathcal{L}^2$ para evitar que se contrarresten termos positivos e negativos.

Unha vez adestrada a rede unha iteración máis (tarefa imprescindible para finalizar a Etapa 2), calculando todos os pesos –incluíndo os da capa de atención das $\mathcal{F}$ variables fixas e das $(\mathcal{T}\backslash\{j_1\})\backslash\mathcal{F}$ variables suxeitas a ser eliminadas–, proporciónase unha nova clasificación (diferente á relativa á Etapa 1, por modificarse os pesos no adestramento) e considérase o novo conxunto de rexistros ben clasificados. Fixando notación, no Paso 3.i, o conxunto de rexistros ben clasificados vén dado por

$$G_i = \begin{cases} \{r \in \{1, ..., n\} : \text{ o rexistro } r \text{ está ben clasificado na Etapa 2}\}, \text{ se } i = 1, \\ \{r \in \{1, ..., n\} : \text{ o rexistro } r \text{ está ben clasificado no Paso 3.} i - 1\}, \text{ se } i > 1. \end{cases}$$

Logo do Paso 3.1, realizado como a primeira repetición da Etapa 2, pero considerando o conxunto $(\mathcal{T}\backslash\{j_1\})\backslash\mathcal{F}$ en lugar do conxunto $\mathcal{T}\backslash\mathcal{F}$, elimínase a variable $j_2 \in (\mathcal{T}\backslash\{j_1\})\backslash\mathcal{F}$ seguindo un razoamento totalmente análogo e, en particular, localizando j_2 tal é como se fixo na obtención de j_1 (segundo a expresión (2.17)). Como se expón no Algoritmo 2, este proceso debe repetirse un número preestablecido de iteracións, finalizando cando se verifique un criterio de parada axeitadamente elixido e que, para o noso caso, determinaremos na subsección adicada aos resultados.

Resultados

O número de variables secuenciais dispoñibles para cada rexistro de ECG ascende a 83 e, como xa se comentou previamente, a súa explicación recóllese no Apéndice B, con especial atención ás variables rítmicas e morfolóxicas. De feito, omitindo os 12 ritmos e as 28 morfoloxías, as variables candidatas ao descarte redúcense a $83 - 12 - 28 = 43$. Por este motivo, e como cota superior do número de variables eliminadas, consideráronse 14 iteracións na Etapa 3, o que ocasiona que, ao sumo, se adestren modelos omitindo $1 + 14 = 15$ variables (unha pola Etapa 2 e 14 pola Etapa 3).

Dado que a ordenación entre os distintos *folds* non coincidía, tivemos que lograr un acordo a través do cálculo dunha medida xeral que permita crear unha ordenación final, incorporando a posición de eliminación. No noso caso, para cada *fold* de validación, ordenáronse as 15 primeiras variables descartadas e, en función diso, definiuse un criterio de acordo para determinar cantas e cales eliminar. Co obxectivo de ilustrar os resultados acadados, no Cadro 2.12 (a) preséntase a ordenación das variables candidatas ao descarte en cada un dos *folds*, indicando a iteración na que foron eliminadas. Como consecuencia, o criterio de parada aplicado pasa por omitir un máximo de 15 variables no adestramento do modelo secuencial para, posteriormente, elixir de forma consensuada aquelas que debemos eliminar.

Nesta situación, débese determinar unha medida de acordo que asigne puntuación nula ás variables nunca eliminadas e puntuación positiva ás variables eliminadas nalgunha ocasión, incorporando a posición de eliminación. Deste xeito, se unha variable nunca se eliminou, a puntuación de eliminación asociada debe ser igual a 0. Ademais, o índice de ordenación é clave na decisión final e é estritamente crecente. Consecuentemente, a medida que o índice de ordenación aumenta, o seu inverso diminúe estritamente. Como a raíz cadrada é unha función monótona crecente en $\mathbb{R}^+$, o inverso da raíz cadrada tamén presenta esta propiedade de monotonía (decrecente).
Segundo o Algoritmo 2, imos considerar a seguinte ecuación, avaliada nos 15 índices de ordenación asociados ás 15 posibles variables eliminadas do modelo:

$$\frac{1}{2}\left(\frac{1}{i+1} + \frac{1}{\sqrt{i+1}}\right), \ i = 0, ..., 14 \text{ iteracións.} \tag{2.18}$$

Fold	Iteracións 0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
fold 0	49	48	52	44	20	35	36	37	30	42	19	16	53	47	43
fold 1	43	30	15	31	23	53	52	38	50	28	12	19	34	49	46
fold 2	31	44	35	29	46	12	52	19	16	36	43	20	33	21	51
fold 3	52	41	39	15	20	12	13	21	36	49	51	16	42	19	44
fold 4	19	30	23	31	46	42	14	12	34	52	29	13	41	48	15
fold 5	34	49	14	42	16	45	28	33	46	51	32	37	26	19	50

(a) Primeiras 15 variables eliminadas en cada *fold*, codificadas segundo a súa posición.

Nº variable1	Puntuación1	Nº variable2	Puntuación2	Nº variable3	Puntuación3	Nº variable4	Puntuación4
12	0.72294439	23	0.7789486	34	1.39935881	45	0.28745748
13	0.44641504	24	0	35	0.74279928	46	0.7082618
14	0.71575261	25	0	36	0.69074691	47	0.16934491
15	0.99277458	26	0.17713659	37	0.42528093	48	0.7728983
16	0.73183325	27	0	38	0.2392767	49	1.98101218
17	0	28	0.46852469	39	0.4553418	50	0.384655
18	0	29	0.57121022	40	0	51	0.56675688
19	1.79083605	30	0.82577561	41	0.78068998	52	1.92386649
20	0.50961103	31	1.375	42	1.04770795	53	0.46459407
21	0.4086216	32	0.19621022	43	1.358643	54	0
22	0	33	0.41641328	44	1.14098617		

(b) Puntuación global das variables en función da métrica dada pola ecuación (2.18).

Cadro 2.12: Resultados principais da selección de variables secuenciais.

Trátase dunha métrica estritamente monótona decrecente en $\mathbb{R}^+$, que pretende equilibrar o peso asignado á cola nun entorno da orixe (pola dereita) e o peso asignado á cola dereita. A idea da súa aplicación é que a posición inflúa na eliminación, pero que non sexa excesivamente determinante.

A continuación, recóllense os principais resultados da selección de variables secuenciais, indicando cales se eliminaron e destacando a importancia daquelas cuxa situación é a oposta.

- Variables nunca eliminadas: 17, 18, 22, 24, 25, 27, 40, 54.
- Variables eliminadas unha soa vez e en pasos avanzados: 47, 26, 32, 38.
- Ordenación das variables segundo a puntuación outorgada pola ecuación (2.18):
  49, 52, 19, 34, 31, 43, 44, 42, 15, 30, 41, 23, 48, 35, 16, 12, 14, 46, 36, 29, 51, 20, 28, 53, 39, 13, 37, 33, 21, 50, 45, 38, 32, 26, 47, 40, 27, 25, 24, 22, 18, 17, 54.
- Variables finalmente eliminadas do modelo: 49, 52, 19, 34, 31, 43, 44, 42 (7 variables).

Para finalizar coa exposición dos resultados, imos realizar algún comentario relacionado co tempo de execución necesario e coa demanda computacional requerida polo Algoritmo 2. Naturalmente, é

claro que esta última é elevada porque debemos adestrar redes neuronais, pero resulta menor en comparación a aplicar validación cruzada a toda a rede, eliminando de cada vez unha variable. Ademais, pódense realizar as operacións para cada variable da mesma iteración en paralelo. Por exemplo, na Etapa 2 (análogo para todos os pasos da Etapa 3) podemos calcular as $|\mathcal{T}\backslash\mathcal{F}|$ diferenzas por separado e logo comparalas para achar cal é a mínima e determinar a variable que debemos eliminar.

En canto ao tempo de execución, na Etapa 1 precísanse 30 minutos se esiximos 40 iteracións iniciais para adestrar á rede. A Etapa 2, e cada paso da Etapa 3, tarda uns 30 segundos en calcular a diferenza para cada variable. En consecuencia, o tempo aproximado é o seguinte:

$$30 \cdot |\mathcal{T}\backslash\mathcal{F}| = 30 \cdot (55 - 12) = 30 \cdot 43 = 1290 \text{ s} = 21.5 \text{ min}.$$

Como o número de variables se vai reducindo recursivamente (eliminando unha variable na Etapa 2 e en cada paso da Etapa 3), en cada iteración o tempo de execución redúcese uns 30 segundos: no Paso 3.1 tarda uns 21 min; no Paso 3.2 uns 20.5 min, etc. Tendo en conta que queremos realizar 15 etapas, obtemos que, para cada *fold* de CV, o Algoritmo 2 tarda

$$30 \text{ min} + \sum_{k=0}^{14} \frac{30}{60} \cdot (43 - k) \text{ min} = (30 + 270) \text{ min} = 5 \text{ h}.$$

Como debemos repetir o proceso nos 6 *folds* de validación, o tempo aproximado total é de 30 horas.

2.2.3. Interpretabilidade da solución

A importancia da interpretabilidade dos resultados non só reside en entender por que un rexistro patolóxico está clasificado como tal –ou achar onde habita o carácter anómalo dunha serie de latexos– senón que permite localizar onde a rede fixa erroneamente a atención para elixir incorrectamente unha etiqueta, así como conforma unha ferramenta de apoio para o persoal sanitario. A implementación de criterios automáticos de diagnóstico e monotorización cardíaca, en ausencia de interpretabilidade, pode clasificar instantaneamente pero non proporciona unha guía en relación aos criterios automáticos de diagnóstico e, polo tanto, non favorece ao modelo de intelixencia artificial en canto a incrementar a súa credibilidade por parte da comunidade médico–científica. No noso caso, e apoiándonos na descrición de ritmos realizada por *Construe*, así como na extracción de variables temporais, o clasificador secuencial dota dunha xustificación propia á elección da etiqueta final, proporcionando un resultado de clasificación transparente e auto–explicable.

Na práctica, a rede concentra a atención en como o conxunto de latexos discrimina e elixe unha clase ou outra, fixándose no trazo completo, que lle dá sentido á distinción de clases, e non na relación entre os diferentes latexos. En consecuencia, a capa de ritmos pode atender aos primeiros e últimos latexos, propoñendo e reafirmando unha certa hipótese, ou pode centrarse en determinados eventos patolóxicos que ocorren en certos fragmentos do rexistro. Por outra parte, a capa xeral atende a aqueles descubrimentos que, pola súa natureza, axudan máis a explicar a elección da clase final. Dende a análise de ritmos presentada por *Construe* ás matrices de pesos obtidas coas redes de atención, o ideal sería que a elección dunha clase ou outra viñese acompañada dunha xustificación gráfica válida, comparable cos criterios adoptados por un posible cardiólogo. A continuación, para ilustrar a interpretabilidade visual da auto–explicabilidade do clasificador secuencial, engádense varios exemplos ilustrativos.

Exemplos ilustrativos

En primeiro lugar, na Figura 2.13 represéntanse, en escala de cor, os pesos asociados á capa de ritmos e á capa xeral de atención para o rexistro A00128, etiquetado e clasificado como **FA**. Coloreando os pesos de atención dos dous camiños que conforman o esquema da rede, podemos indagar que

latexos son importantes para a clasificación e cales son irrelevantes, detectando onde se sitúa a información latente no rexistro de entrada. En concreto, esta referencia ten etiquetados todos os latexos como de FA (ritmo 10 na codificación de *Construe*, **Rh**), trasladándose isto a unha capa rítmica con pesos equilibrados, agás os primeiros e os últimos latexos, que caracterizan e confirman a fibrilación auricular. Na capa xeral, latexos cun patrón de arritmia máis marcado teñen asociados pesos lixeiramente superiores. A probabilidade que proporcionan ás redes á clase **FA** é 0.97, 10 décimas superior á probabilidade que proporciona o XGBoost, que é 0.87. Precisamente, o rexistro A00128 é un exemplo no cal ambos clasificadores están seguros da etiqueta final, que resulta ser a acertada.

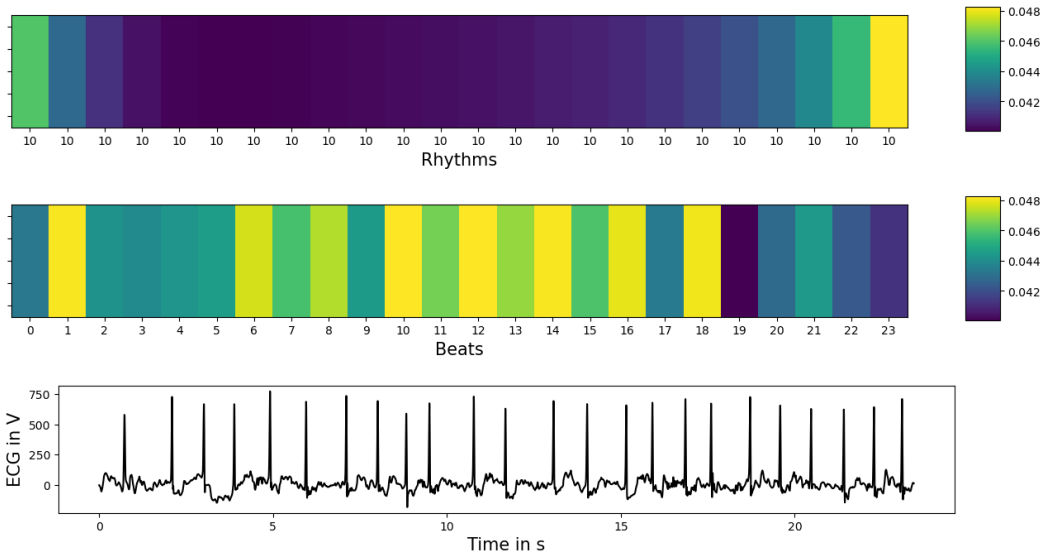
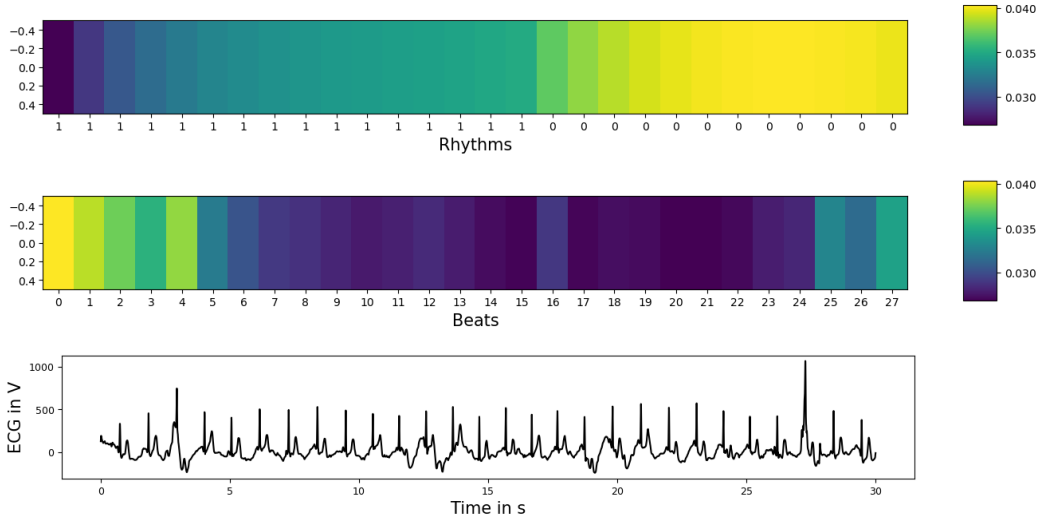


Figura 2.13: Rexistro A00128 (**FA**): gráfica secuencial e vectores de atención.

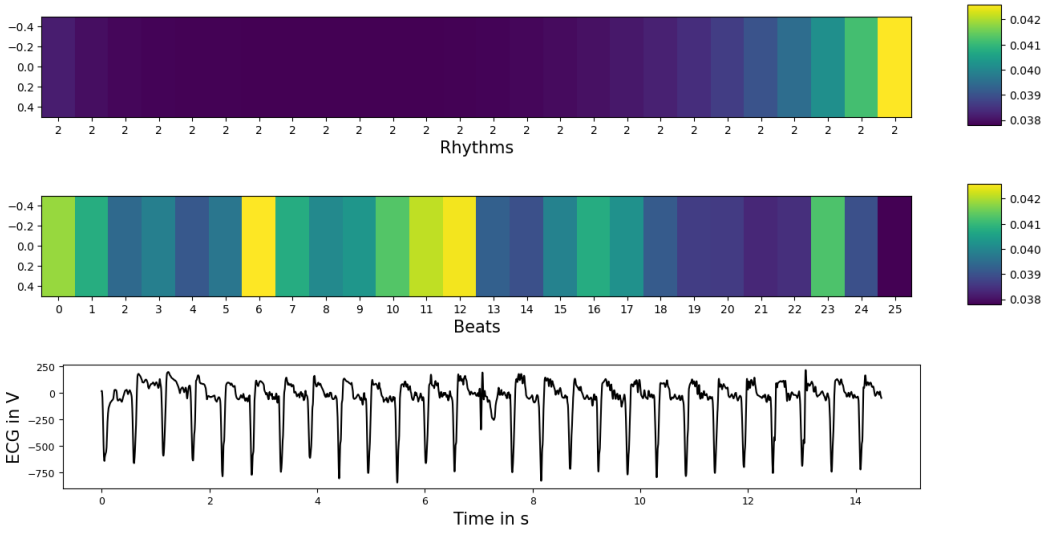
Para terminar con esta sección, inclúense tres comentarios adicionais, novamente relacionados coa interpretabilidade das redes de atención, seguindo un razoamento similar ao aplicado para o rexistro A00128 (**FA**). Neste caso, ímonos centrar nun exemplo de normalidade e en dous doutras patoloxías, englobadas baixo a etiqueta **O**. Na Figura 2.14 represéntanse as gráficas secuenciais e os vectores de atención obtidos ao procesar os rexistros A00104 (**N**), A02434 (**O**) e A00123 (**O**).

Como podemos observar na Figura 2.14 (a), o rexistro A00104 presenta bradicardia na primeira metade, pero sen ser suficiente como para encadralo como anómalo, centrándose a capa de ritmos na metade restante, correspondente a ritmo sinusal. A referencia está gardada coa etiqueta **N**, común a súa clasificación, con 58 pulsacións/minuto, lixeiramente inferior ao limiar de bradicardia. Fixándonos na gráfica secuencial, semella que todos os latexos se atopan en ritmo e presentan unha morfoloxía persistente e non patolóxica, centrándose a capa xeral nos primeiros e últimos latexos que, similares aos restantes, confirman a hipótese de normalidade. Tanto as redes de atención –cunha probabilidade de 0.9937– como o XGBoost –cunha probabilidade de 0.9947– propoñen con firmeza a clase **N**.

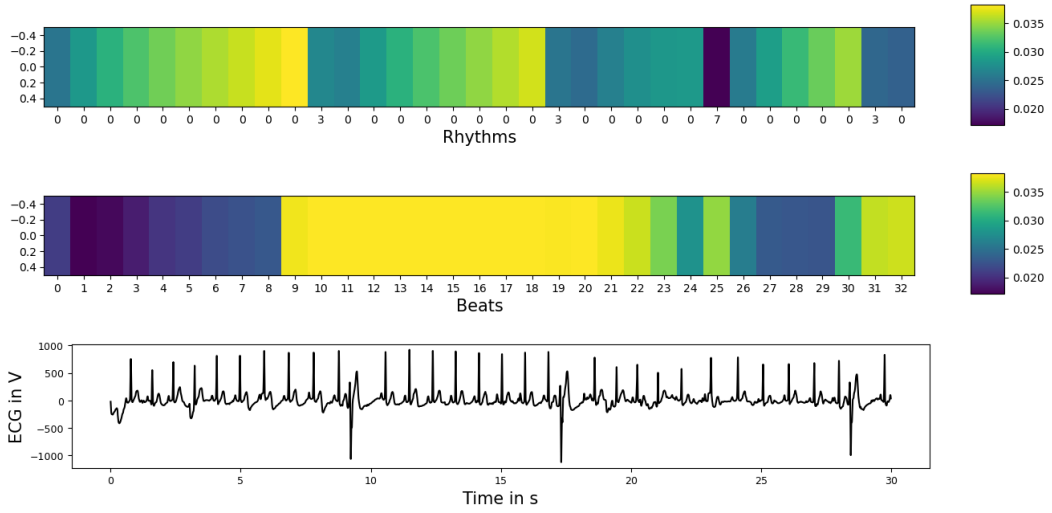
Por outra banda, segundo a Figura 2.14 (b), todos os latexos do rexistro A02434 son taquicárdicos, conformando unha evidencia que o etiqueta e clasifica como patolóxico, cunha probabilidade de 0.9705 das redes e de 0.9974 do XGBoost. A capa de ritmos asigna maior peso aos primeiros e, sobre todo, aos últimos latexos, propoñendo e afirmando a hipótese de taquicardia, focalizando a súa atención a capa xeral no número 13, probablemente polo cambio de morfoloxía e amplitude. Trátase dun rexistro invertido pero isto non supón un problema: *Construe* comproba previamente a posible inversión e, en caso afirmativo, revértea. Finalmente, examinando a Figura 2.14 (c), o rexistro A00123 presenta



(a) A00104 (**N**): bradicardia inestimable en beneficio da etiqueta de normalidade.



(b) A02434 (**O**): inversión do rexistro e taquicardia que persiste no tempo.



(c) A00123 (**O**): multitude de latexos ectópicos (ventriculares), con saltos de amplitude.

Figura 2.14: Gráfica secuencial e vectores de atención de varios rexistros ben clasificados.

tres latexos ventriculares, con morfoloxía claramente distinta á dos demais e que, en todos os casos, se detectan erroneamente como extrasístoles (como consecuencia dun lixeiro adianto na aparición do complexo QRS). A pesar deste erro, o seu carácter anómalo conforma unha evidencia suficiente como para clasificar o rexistro como patolóxico. Por último, detéctase incorrectamente un bloqueo de rama a causa dun pequeno cambio de ritmo, pero a capa de atención asígnalle peso ínfimo en comparación ao dos outros latexos. Tanto na capa de ritmos como na xeral, os pesos concéntranse próximos aos latexos ventriculares, o que explica a etiqueta final de **O**, cunha probabilidade de 0.9865 das redes e de 0.9833 do XGBoost.

En definitiva, o anterior estudo gráfico, relativo á interpretabilidade do clasificador secuencial, pódese estender a calquera rexistro da base de datos –e incluso a rexistros externos á mesma– por ser unha análise xeral que proporciona unha solución universal. Como resultado, o clasificador secuencial permite localizar onde a rede fixa a atención para elixir unha etiqueta ou outra, proporcionando un resultado auto–explicable tanto para unha actuación correcta como para detectar que latexos motivan a incorrecta clasificación dun rexistro. Isto permítenos obter unha primeira idea das propiedades morfolóxicas/rítmicas comúns ás referencias mal clasificadas e extraer patróns ordinarios, axudando a entender os erros cometidos. No Capítulo 4 listaranse os problemas de interpretabilidade achados e diferentes intentos de mellora que, por diversos motivos, non lograron o resultado esperado.

2.3. Clasificación *ensemble*

Os métodos combinados ou de *ensemble* (Wolpert, 1992) son modelos de nivel superior, é dicir, cuxas variables de entrada son as probabilidades de saída doutros modelos, denominados de nivel inferior. O seu fundamento radica en combinar as predicións dos modelos de nivel inferior para mellorar o rendemento e xeralización individual. No noso caso, como metaclasificador aplicado na etapa de *ensemble*, onde usamos as probabilidades de saída do clasificador global e secuencial como variables de entrada (eliminando as asociadas á clase ∽, por non tratarse dun ritmo cardíaco, senón dun ritmo ruidoso, inclasificable pola súa baixa calidade) empregamos unha máquina de soporte vectorial (*support vector machine*, SVM) con núcleo lineal[13]. A idea é buscar o hiperplano separador de máxima marxe, que minimice o erro de clasificación lineal, para proporcionar unha etiqueta final que perfeccione o desempeño individual dos clasificadores de nivel inferior, manexando un hiperparámetro de regularización que relaxe as restricións de clasificación lineal para evitar sobreaxuste.

Axudándonos do contido exposto no Capítulo 7 de Bishop (2006), vamos presentar as SVM, estudando o cálculo dos parámetros mediante a resolución dun problema de optimización convexa (que garante que todo óptimo local é global) con $K > 2$ clases. Ante esta problemática, dado que a metodoloxía SVM está especialmente deseñada para a clasificación binaria, e o noso interese é a multiclase, comezamos presentando os dous enfoques máis populares para a súa extensión a $K > 2$ clases, combinando SVM binarias co obxectivo de construír un clasificador multiclase:

1. Un contra un (*one versus one*, OVO): constrúense $K(K-1)/2$ SVM binarias cos posibles pares de clases, clasificando unha nova observación de acordo á clase máis votada entre todos os modelos.

   No noso caso, con 6 SVM binarias, escolleríase a clase máis votada entre $\{\mathbf{N}, \mathbf{FA}, \mathbf{O}, \backsim\}$.

2. Un contra o resto (*one versus the rest*, OVR): constrúense K SVM binarias, onde o k-ésimo modelo se adestra usando os datos da clase k por un lado e os datos das restantes $K-1$ agrupados. Como resultado, obtemos un problema de clasificación binaria por cada clase.

   No noso caso, con 4 SVM binarias, escolleríase a clase máis acertada entre $\{\mathbf{N}, \mathbf{FA}, \mathbf{O}, \backsim\}$. Por exemplo, se a clase final proporcionada fose **FA**, obteríamos o seguinte resultado:

[13] Cun núcleo lineal, a SVM é equivalente ao clasificador de soporte vectorial (*support vector classifier*).

- Clasificador de **N** fronte ao resto: a clase do rexistro observado non é **N**.
- Clasificador de **FA** fronte ao resto: a clase do rexistro observado é **FA**.
- Clasificador de **O** fronte ao resto: a clase do rexistro observado non é **O**.
- Clasificador de **ᔓ** fronte ao resto: a clase do rexistro observado non é **ᔓ**.

Posto que o enfoque OVR non é axeitado cando as clases están desbalanceadas –como é o noso caso– aplicaremos o enfoque OVO, que ten como inconveniente que para valores grandes de K é preciso adestrar moitos clasificadores. Sen embargo, para $K = 4$ clases só serán necesarios 6 modelos diferentes. Esta discusión xustifica o uso da opinión maioritaria como vía de extensión a $K > 2$ das SVM e fundamenta a exposición da formulación matemática dos modelos binarios.

O obxectivo é calcular o hiperplano separador de máxima marxe (*maximal margin classifier*), que é o que minimiza o erro de clasificación lineal, onde se define a marxe como a menor distancia posible entre dous vectores de variables procedentes de categorías diferentes. O nome do método débese aos vectores de soporte, que son as observacións que menos distan do hiperplano de máxima marxe (soportan en equilibrio a este hiperplano de separación, que é aquel que minimiza o erro de clasificación lineal). No caso das SVM, inclúese un parámetro C que relaxa a clasificación lineal (a suposición de separabilidade dos datos non é frecuente na práctica) e engade robustez (penaliza por mala clasificación), e unha función núcleo K, dada por:

$$K(\cdot, *) = \phi(\cdot)^t \phi(*), \tag{2.19}$$

que se aplica ás variables explicativas para conseguir unha relación lineal no espazo transformado, (comunmente) de dimensión superior, que se traduce en regras de clasificación non lineais no espazo orixinal. No Apéndice C.3 defínense as funcións núcleo máis comúns e detállase a nosa elección.

Chegados a este punto, consideremos que dispoñemos dun conxunto de adestramento

$$\{(\overrightarrow{x_i}, y_i) \in \mathbb{R}^D \times \{-1,\ +1\}\}_{i=1}^n,$$

onde $\overrightarrow{x_i} = (x_{i1}, ..., x_{iD}) \in \mathbb{R}^D$ é un vector de observacións das D variables explicativas, que constitúen o vector aleatorio $\overrightarrow{X} = (X_1, ..., X_D)$, e y_i é a i-ésima observación da variable resposta Y que, por conveniencia, establecemos que teña como categorías o -1 e o $+1$ (porque será o signo o que determine a clase dunha nova observación).

En concreto, quérese calcular un vector $\omega \in \mathbb{R}^D$ e un escalar $b \in \mathbb{R}$ tal que a clase dada por

$$\text{signo}(\omega^t \phi(\overrightarrow{X}) + b)$$

sexa correcta para a maioría das observacións, onde o signo positivo se asocia á categoría $+1$ e o negativo á categoría -1. Nesta formulación, o nesgo (b) considérase independente do vector de pesos (ω) para facilitar a interpretabilidade do hiperplano clasificador (a diferencia da notación das ANN). O problema de optimización a resolver vén dado por

$$\begin{aligned} \text{(SVM–primal)}\ \min_{\omega, b, \zeta} & \left(\frac{1}{2}\omega^t\omega + C\sum_{i=1}^{n} \zeta_i \right) \\ \text{s.a} \quad & y_i(\omega^t \phi(\overrightarrow{x_i}) + b) \geqslant 1 - \zeta_i,\ i = 1, ..., n \\ & \zeta_i \geqslant 0,\ i = 1, ..., n, \end{aligned}$$

onde o termo $\omega^t\omega$ da función obxectivo se incorpora para maximizar a marxe, e co segundo sumando se engade unha penalización que actúa cando unha mostra está mal clasificada (*soft margin classifiers*). Unha clasificación perfecta implicaría que

$$y_i(\omega^t \phi(\overrightarrow{x_i}) + b) \geqslant 1,\ \forall i = 1, ..., n.$$

Como é habitual que non sexa posible (sen sobreaxustar), engádense n variables de holgura ζ_i que modelan a distancia das observacións á fronteira de separación, controladas polo hiperparámetro C. A idea é maximizar a marxe á vez que penalizamos os puntos que se atopan no lado equivocado. Por último, ϕ é unha transformación que se aplica ás variables de entrada para incrementar a flexibilidade e que está directamente relacionada coa definición de función núcleo dada pola expresión (2.19).

Para a resolución do problema (SVM–primal), formúlase a súa versión dual, dada por

$$\begin{aligned} \text{(SVM–dual)} \quad & \min_{\alpha} \left(\frac{1}{2} \alpha^t Q \alpha - \mathbb{1}_n \right) \\ & \text{s.a} \quad \overrightarrow{y}^t \alpha = 0, \\ & \qquad 0 \leqslant \alpha_i \leqslant C, \ i = 1, ..., n, \end{aligned}$$

onde $Q \in \mathcal{M}_{n \times n}$ é unha matriz semidefinida positiva tal que as súas entradas veñen dadas por

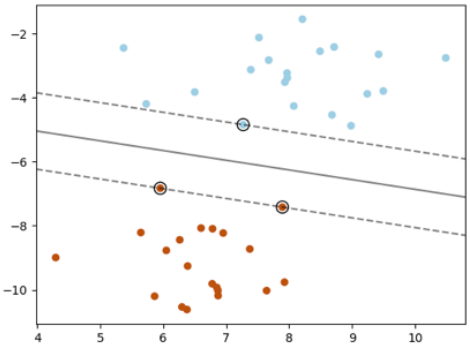
$$Q_{ij} = y_i y_j K(\overrightarrow{x_i}, \overrightarrow{x_j}),$$

$\mathbb{1}_n$ é un vector $n-$dimensional con todas as entradas unitarias, $\overrightarrow{y} = (y_1, ..., y_n) \in \mathbb{R}^n$ é un vector cuxas compoñentes se corresponden coas observacións da variables resposta Y e $\alpha \in \mathbb{R}^n$ é o vector de variables duais. Finalmente, a clase dunha nova observación $(\overrightarrow{x})$ vén determinada por

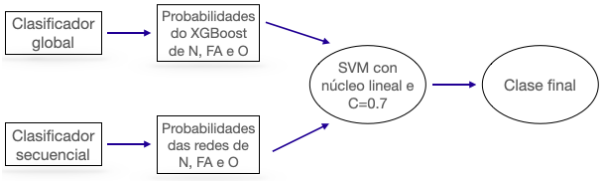
$$\text{signo} \left(\sum_{i \in SV} y_i \alpha_i K(\overrightarrow{x_i}, \overrightarrow{x}) + b \right),$$

onde só precisamos sumar sobre os índices relativos ao conxunto de vectores de soporte, nomeado como SV, porque α_i será igual a 0 para o resto de observacións.

Na Figura 2.15 (a) represéntase un hiperplano de máxima marxe para unha clasificación binaria, onde a cor das observacións determina a súa clase. Neste exemplo, obtéñense 3 vectores de soporte, proporcionando unha separación lineal entre as observacións do conxunto de adestramento perfecta. Por outra banda, na Figura 2.15 (b) recóllese a arquitectura da SVM implementada, onde a elección do núcleo e a optimización do hiperparámetro C se discute no Apéndice C.3.



(a) Exemplo de hiperplano de máxima marxe, con 3 vectores de soporte e separación lineal das clases.



(b) Arquitectura da SVM implementada.

Figura 2.15: Base dunha SVM binaria e particularización para a resolución do noso problema.

En relación ao modelo exposto en Teijeiro et al. (2018b), substitúese a análise lineal discriminante (*linear discriminant analysis*, LDA) por unha SVM, dado que non se cumpren as hipóteses de optimalidade do primeiro metaclasificador, sendo máis conveniente utilizar un menos restritivo: o LDA é máis

simple e funciona mellor baixo normalidade e igualdade de varianzas e matrices de covarianzas pero, se isto non se cumpre, os resultados que proporciona poden ser inconsistentes. É sinxelo comprobar –por exemplo cun test de normalidade de Shapiro-Wilk univariante– que ningunha das poboacións (entendendo as distintas poboacións como os $3 \cdot 2 = 6$ grupos de probabilidades proporcionadas polo XGBoost e polas redes) se distribúe segundo unha normal univariante, resultando que o vector aleatorio 6-dimensional do que se obteñen as variables de entrada do LDA non segue unha normal multivariante[14]. Consecuentemente, tampouco é aconsellable utilizar a análise cadrática discriminante (*quadratic discriminant analisys*, QDA), que tamén esixe normalidade do vector de entrada. Por outra banda, poderíamos usar un clasificador que implemente un método dos veciños máis próximos (*K-nearest neighbors classifier*, KNN), establecendo previamente o número de veciños e os pesos dos mesmos. Sen embargo, logo de realizar as correspondentes probas comparativas, obtivemos mellores resultados coas SVM que cos KNN, xustificando descartar o uso destes últimos.

Unha vez rematado o adestramento e optimización dos tres clasificadores utilizados no presente traballo, engadíronse varios criterios de intervención externos á aprendizaxe estatística, fundamentados na natureza do problema e no significado das clases, definidas segundo as diferentes arritmias máis comúns. O obxectivo é perfeccionar a clasificación final, modificando automaticamente a etiqueta de certos rexistros, determinados por unha serie de pautas que se recollen a continuación:

(a) Se XGBoost clasifica ∽, a etiqueta final será ∽.

(b) Se a rede clasifica ∽, a etiqueta final só utilizará información das probabilidades do XGBoost.

(c) Se XGBoost clasifica **FA** cunha probabilidade superior ao 50 %, a etiqueta final será **FA**.

Este conxunto de intervencións externas ten uns fundamentos moi intuitivos. Por unha parte, a finalidade de (a)-(b) é dotar ao clasificador global da decisión de etiquetar un rexistro como ruidoso, por ser o ruído unha propiedade máis global que local, difícil de detectar latexo a latexo pero facilmente identificable en agregado. Por outra parte, a finalidade de (c) é diagnosticar FA se o modelo global ten evidencias suficientes que sosteñan esta decisión, ao caracterizarse como un ritmo cardíaco irregular e non constante, cunha distancia temporal entre latexos desigual. No Capítulo 3 presentaranse os resultados estatísticos do modelo proposto e, en concreto, do clasificador *ensemble* modificado. Sen embargo, para entender as vantaxes que supón engadir os anteriores criterios, imos adiantar unha serie de resultados. Concretamente, e en relación aos valores acadados nas matrices das Táboas 3.3 e 3.4 e –aínda que se omiten por carecer de importancia– tendo en conta os relativos ao clasificador *ensemble* posterior ás modificacións (a) e (b) (sen engadir a modificación (c)), obtemos que:

- Os criterios (a) e (b) melloran a clasificación de ∽ e non teñen efecto na correcta clasificación de rexistros etiquetados como **FA** e **O**. Non obstante, empeoran a clasificación de **N** pero, en base a proporción mostral de **N** e ∽, priorízase a discriminación de ∽.
- O criterio (c) ten un resultado positivo en canto á correcta clasificación de **FA**, mellorando os rexistros confundidos con **O**. Comparado co escenario anterior, non hai cambios relevantes en relación aos resultados de clasificación de **N** e ∽.

[14] Recordemos a seguinte propiedade da distribución normal multivariante: Calquera subconxunto de variables dun vector $\overrightarrow{X} = (X_1, ..., X_D)$ con distribución normal $d-$dimensional, $N_D(\mu, \Sigma)$, ten distribución normal (con vector de medias e matriz de covarianzas correspondentes á partición de $\overrightarrow{X}$). En particular, cada compoñente X_i ten distribución normal univariante, denotada por $N(\mu_i, \sigma_i^2)$. Sen embargo, sen engadir máis restricións, o recíproco non é certo. Para que un vector aleatorio constituído por variables aleatorias normais univariantes sexa normal multivariante, as variables aleatorias que o conforman deben ser mutuamente independentes.

Capítulo 3

Resultados estatísticos

Neste capítulo descríbese en detalle a tarefa de validación do modelo proposto, destinado á clasificación (supervisada) de rexistros de ECG, segundo a súa clase corresponda con **N**, **FA**, **O** ou ∽. Para este cometido, preséntanse distintas medidas de calidade da clasificación, discútese a súa funcionalidade na aplicación de interese e amósanse os resultados numéricos e gráficos obtidos. Acto seguido, abórdase o caso binario, asociado ao cribado de FA, e calcúlanse métricas derivadas deste clasificador.

Para a tarefa de validación, o máis habitual na inferencia estatística clásica é utilizar toda a información dispoñible (todos os datos) para construír un modelo e, asumindo a súa validez, empregar métodos inferenciais que midan a súa precisión e bo funcionamento. Sen embargo, é común que os modelos de aprendizaxe estatística estean hiperparametrizados, ocasionando esta excesiva flexibilidade posibles problemas de sobreaxuste e escasa xeralización, obtendo un clasificador moi a medida para as observacións coñecidas no adestramento pero que, potencialmente, non terá un bo comportamento ao aplicarse a novas entradas. Ante esta problemática, se non se dispón dun conxunto de test, o método máis simple para estimar o erro de clasificación é, probablemente, a validación cruzada. Por este motivo, co fin de coñecer como de xeralizable é o modelo proposto, utilizamos validación cruzada de k pregaduras (k-*fold* CV, *k-fold cross validation*), dividindo os datos en k subconxuntos estratificados. Para cada posible combinación, úsanse $k-1$ conxuntos no adestramento e o restante é sobre o que se calculan as medidas de calidade e se estuda o rendemento do clasificador. O resultado final é o promedio destas medidas calculadas en todas as combinacións.

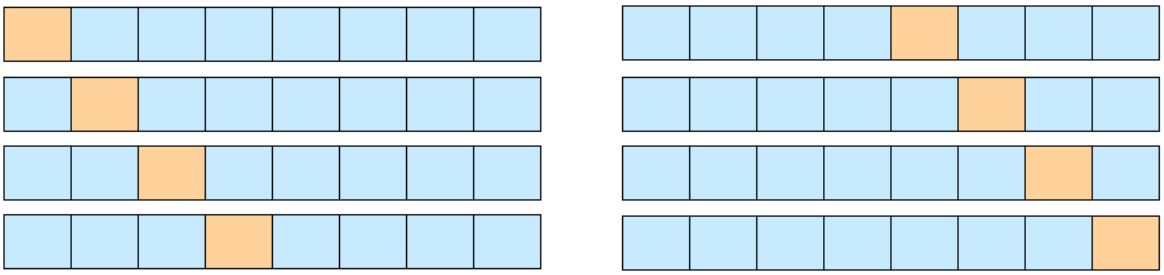


Figura 3.1: Esquema de 8-*fold* CV: En laranxa, o conxunto de validación e en azul, os de adestramento.

Neste escenario de validación, obtemos que para $k = n$ se define a CV clásica, que de cada vez deixa un dato fóra (*leave-one-out* CV), obtendo o método de CV máis exhaustivo pero cun custo computacional excesivo e aplicabilidade limitada. No lado oposto, para $k = 1$ atopámonos ante un caso patolóxico inaplicable, onde ningún dato se destina ao adestramento ($k-1=0$) e todos se reservan para a validación, conformando unha situación externa á CV onde, necesariamente, se debe esixir que $k > 1$ (loxicamente, $k \leqslant n$). No noso caso, co fin de obter un equilibrio exhaustividade–rendemento,

aplicamos 8-*fold* CV, cuxo esquema de división se representa na Figura 3.1.

En canto á porcentaxe das etiquetas, o 64.02 % son **N**, o 8.60 % son **FA**, o 22.98 % son **O** e o 3.62 % restante correspóndese a ruído. Este significativo desequilibrio na procedencia dos rexistros da base de datos ocasiona que medidas de calidade da clasificación como a exactitude (*acuraccy*) –entendida como a proporción de rexistros ben clasificados– non sexan axeitadas por non ter en conta a distribución das clases. Segundo estas porcentaxes, a taxa de acertos non é un bo indicador porque non manexa con igual importancia a correcta discriminación de todas as clases, presentando nesgos prexudiciais cara a máis representada, conseguindo solucionar este problema co uso da medida $F1$.

Na presentación do *Physionet*/CinC *Challenge* 2017 exposta no Prefacio establécese que a etiqueta de ruído se asocia a rexistros de moi baixa calidade e, consecuentemente, resulta lóxico excluíla da análise do rendemento dos clasificadores. Así mesmo, defínese a métrica $F1$ por ser a medida oficial de puntuación do reto. Recordando a súa expresión, calculada como a media aritmética das medidas $F1$ das restantes clases, determinamos que a medida $F1$ multiclase final se define como:

$$F1 = \frac{F_{1N} + F_{1FA} + F_{1O}}{3}, \tag{3.1}$$

sendo F_{1L} a medida $F1$ asociada á clase $\mathbf{L} \in \{\mathbf{N}, \mathbf{FA}, \mathbf{O}\}$, tal é como se expón no Prefacio.

Co obxectivo de visualizar o desempeño do clasificador final, obtendo unha idea da magnitude das equivocacións entre dúas clases, usase a matriz de confusión (*confusion matrix*), que é unha táboa de continxencia onde cada columna representa o número de predicións de cada clase, mentres que cada fila representa o número de etiquetas da clase real. No noso caso, e para a discriminación de catro clases (**N**, **FA**, **O** e ∽), a súa forma vén dada pola Táboa 3.1, onde na diagonal se recollen a cantidade de rexistros ben clasificados (*true class*, TC) e nas outras entradas se indican as correspondentes equivocacións na asignación de etiquetas.

Observado/ Clasificado	**N**	**FA**	**O**	∽
N	TC			
FA		TC		
O			TC	
∽				TC

Táboa 3.1. Estrutura da matriz de confusión no eido do *Physionet*/CinC *Challenge* 2017.

Observado/ Clasificado	**FA**	Resto
FA	VP: Verdadeiros positivos	FN: Falsos negativos
Resto	FP: Falsos positivos	VN: Verdadeiros negativos

Táboa 3.2. Estrutura da matriz de confusión para o caso binario, relativo ao cribado de FA.

Centrándonos no cribado de FA, particularízase a clasificación ao caso binario, asociado a considerar a clase **FA** e unha nova clase auxiliar, formada pola unión das restantes. Nesta situación, poderíamos cuantificar os falsos positivos (FP), os falsos negativos (FN), os verdadeiros positivos (VP) e os verdadeiros negativos (VN), entendendo como positivo a valoración confirmativa dun rexistro de padecer FA. Entre as métricas binarias típicas, calculouse a sensibilidade (Sens ou Rec: proporción de suxeitos enfermos de FA ben clasificados) –tamén coñecida como taxa de verdadeiros positivos e que mide a capacidade do modelo para identificar a un paciente enfermo– e a especificidade (Esp: proporción de suxeitos non enfermos de FA ben clasificados) –tamén coñecida como taxa de verdadeiros negativos e

que mide a capacidade do modelo para identificar a un paciente san–, definidas como

$$\text{Rec} = \frac{\text{VP}}{\text{VP} + \text{FN}} \quad \text{e} \quad \text{Esp} = \frac{\text{VN}}{\text{VN} + \text{FP}}.$$

Na Táboa 3.2 recóllese a estrutura básica da matriz de confusión asociada ao caso binario que, en particular, ten como entradas os indicadores que acabamos de presentar (VP, FN, FP e VN).

A continuación, amósanse os resultados das métricas de desempeño obtidas aplicando 8-*fold* CV. Empezando coa matriz de confusión asociada ao uso directo do clasificador final (Táboa 3.3), observamos que as súas entradas melloran notoriamente (en conxunto) ao engadir os criterios de intervención comentados na Sección 3 do Capítulo 2 (Táboa 3.4). Como se esperaba, estas pautas externas favorecen a clasificación de **FA** e ∽, prexudicando a de **N** e con escaso efecto na de **O**.

Observado/ Clasificado	**N**	**FA**	**O**	∽
N	5464	4	82	27
FA	7	707	29	4
O	45	22	1845	0
∽	9	0	3	278

Táboa 3.3. Clasificador *ensemble*.

Observado/ Clasificado	**N**	**FA**	**O**	∽
N	5446	4	81	17
FA	8	713	31	1
O	45	16	1843	0
∽	26	0	4	291

Táboa 3.4. Clasificador *ensemble* modificado.

Centrándonos na Táboa 3.4, a maior porcentaxe de equivocacións atópase entre **N** e **O**, con 45 rexistros normais clasificados como patolóxicos e 81 patolóxicos na situación contraria. Na Sección 4.1, relativa ás conclusións experimentais, enuméranse unha serie de obstáculos descubertos na etiquetaxe da base de datos (por culpa dos anotadores) ou ocasionados por erros internos na definición de *Construe*, e que impiden mellorar os resultados presentados. Finalmente, tamén existe unha porcentaxe indesexablemente alta de rexistros doutras patolóxicos confundidos con FA e viceversa.

Method | Fold Number	1	2	3	4	5	6	7	8	Mean	SD
XGBoost	0.8733	0.8724	0.8548	0.8601	0.8576	0.8839	0.8674	0.8634	0.8666	0.008
Attention Neural Networks	0.8919	0.907	0.9112	0.8778	0.9128	0.8908	0.9088	0.8976	0.8997	0.0103
SVM-Modified-Stacking	0.9093	0.9127	0.9007	0.9059	0.9097	0.9159	0.9008	0.9083	0.9079	0.0044

Cadro 3.2: Resultados do modelo proposto e dos 2 clasificadores de nivel inferior. Medidas $F1$, media e desviación típica nos 8 *folds* de CV estratificada.

No Cadro 3.2 recóllense os resultados do noso modelo para os 8 *folds* de CV, xunto coa media e a desviación típica. Analizando estas cantidades, concluímos que –en termos de $F1$– a técnica de *ensemble* modificada mellora o desempeño individual dos clasificadores de nivel 1, aumentando a media e sen incrementos preocupantes en canto á desviación típica, que duplica á do XGBoost, pero segue sendo ínfima. Ademais, a medida $F1$ é superior a 0.90 para os 8 *folds*, o que, en certa medida, garante

que esta métrica supera a ese valor, cunha media sumamente prometedora (moi próxima a 0.91).

Na Táboa 3.5 (modelo binario sen intervir) e na Táboa 3.6 (modelo binario intervido), preséntanse as matrices de confusión asociadas ao caso binario, é dicir, ao problema de cribado de FA.

Observado/ Clasificado	**FA**	Resto
FA	707	40
Resto	26	7753

Táboa 3.5. Clasificador binario.

Observado/ Clasificado	**FA**	Resto
FA	713	40
Resto	20	7753

Táboa 3.6. Clasificador binario intervido.

Fixándonos nos resultados de cribado finais (Táboa 3.6), hai 713 referencias correctamente clasificadas como **FA**, isto é, a cantidade mostral de verdadeiros positivos é VP=713. Por outra banda, hai 7753 referencias clasificadas correctamente como non **FA**: a cantidade mostral de verdadeiros negativos é VN=7753. Ademais, os falsos negativos son FN=40 e os falsos positivos ascenden a FP=20, cun total de $\mathrm{FN} + \mathrm{FP} = 40 + 20 = 60$ referencias mal clasificadas, resultado verdadeiramente baixo en comparación coas $\mathrm{VP} + \mathrm{VN} = 713 + 7753 = 8466$ referencias ben clasificadas. Os criterios de intervención externos corrixen a clasificación de 6 rexistros procedentes de pacientes sans.

Na Figura 3.3 trázanse 8 curvas ROC (*receiver operating characteristic*) para o clasificador final, relativas a cada *fold* de adestramento–validación. Trátase de representacións gráficas de un menos a especificidade fronte á sensibilidade do clasificador binario de FA, segundo o limiar de discriminación. Cambiando a terminoloxía, tamén se poden ver como a taxa de falsos positivos (*false positive rate*, FPR), dada por $\mathrm{FPR} = 1 - \mathrm{Esp}$, confrontada coa taxa de verdadeiros positivos (*true positive rate*, TPR), coincidente coa sensibilidade ($\mathrm{TPR} = \mathrm{Rec}$). O modelo de predición óptimo situaríase na coordenada $(0, 1)$ –esquina superior esquerda da gráfica– cun valor de especificidade e sensibilidade de 1 (clasificación perfecta). A recta diagonal que une os puntos $(0, 0)$ e $(1, 1)$ correspóndese co esperado para un clasificador aleatorio, que clasifica **FA** e non **FA** arbitrariamente cunha probabilidade de 0.5. Neste caso, e como se desexa para fundamentar a validez do modelo proposto, o comportamento é moi semellante en todos os *folds* (o reparto dos rexistros foi aleatorio e equidistribuído nas clases) e as curvas toman valores próximos ao punto $(0, 1)$, con áreas baixo elas (*area under the ROC curve*, *AUC*) cercanas a 0.99. Consecuentemente, resulta que a probabilidade de clasificar correctamente dous rexistros elixidos ao azar, un con diagnóstico confirmativo de FA e outro sen el, é próxima ao 99 %.

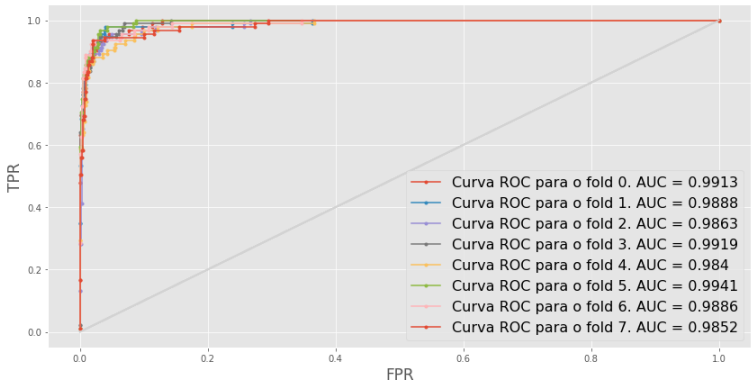


Figura 3.3: Curvas ROC e valores AUC nos 8 *folds* de validación cruzada.

No Cadro 3.4 amósanse os valores de sensibilidade, especificidade, $F1$ e AUC obtidos en cada *fold* de validación e relativos ao clasificador binario. Observamos que, de media, a sensibilidade é cercana a 0.83 e a especificidade a 0.99, reducíndose a medida $F1$ de case 0.91 para $\{\mathbf{N}, \mathbf{FA}, \mathbf{O}\}$ a 0.8534 para **FA**, xustificando unha vez máis a complexidade do problema de cribado. Sen embargo, a desviación típica relativa aos valores $F1$ dos 8 *folds* é de 0.0065, superior ao valor de 0.0044 proporcionado nas mesmas circunstancias para o clasificador multiclase o que, en certa medida, indica o incremento da variabilidade do rendemento do clasificador binario nos distintos *folds* e permite afirmar que o valor $F1$ promedio é menos variable para o modelo multiclase.

FA Measure	1	2	3	4	5	6	7	8	Mean	SD
Sensitivity	0.7926	0.7935	0.8587	0.8478	0.8152	0.8571	0.7912	0.8681	0.828	0.0277
Specificity	0.9938	0.9928	0.9836	0.9846	0.9897	0.9887	0.9959	0.9846	0.9892	0.0039
F1	0.8471	0.8488	0.8449	0.8562	0.8475	0.8667	0.8623	0.8541	0.8534	0.0065
AUC	0.9913	0.9888	0.9863	0.9919	0.9841	0.9941	0.9886	0.9851	0.9888	0.0029

Cadro 3.4: Métricas relativas á adaptación do clasificador final ao caso binario, destinado ao cribado de FA, nos 8 *folds* de validación cruzada, xunto coas medias e desviacións típicas.

Retomando a descrición das medidas de calidade, a taxa de acertos (*accuracy*, ACC) defínese como

$$ACC = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{VP} + \mathrm{VN} + \mathrm{FP} + \mathrm{FN}}$$

e proporciona unha medida de precisión global do clasificador binario (xeralizable facilmente a K clases). Sen embargo, como xa comentamos anteriormente, se as clases non están balanceadas, como é o noso caso, con 733 rexistros etiquetados como **FA** e 7793 etiquetados como non **FA**, esta medida non é oportuna. Ante esta problemática, ademais do cálculo da métrica $F1$, é habitual manexar outras medidas de precisión global que tratan de solventar o desequilibrio no reparto das etiquetas. Por exemplo, a taxa de acertos balanceada (*balanced accuracy*, BA) defínese como

$$BA = \frac{\mathrm{Rec} + \mathrm{Esp}}{2} = \frac{1}{2}\left(\frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}} + \frac{\mathrm{VN}}{\mathrm{VN} + \mathrm{FP}}\right).$$

No noso caso, substituíndo os valores medios de sensibilidade e especificidade, obtemos que

$$ACC = \frac{8466}{8526} = 0.99230 \quad \text{e} \quad BA = \frac{0.828 + 0.9892}{2} = 0.9086,$$

sendo a exactitude balanceada unha medida de calidade que acada un valor menor pero máis axeitado.

Finalmente, en relación aos resultados do clasificador binario, debemos ter en conta que unha elevada porcentaxe de falsos positivos implicaría un importante malgaste de recursos en pacientes sans, incrementando innecesariamente o custe sanitario (e saturación hospitalaria asociada) e prexudicando a rentabilidade do modelo como mecanismo de cribado automático de FA. Sen embargo, unha elevada porcentaxe de falsos negativos supoñería o desacertado diagnóstico de moitos pacientes enfermos –considerando que están sans ou padecen outra doenza cardíaca– o que posteriormente desencadearía consecuencias peores, froito do agravamento dos síntomas da FA. Neste sentido, os procesos diagnósticos manexan probabilidades que axudan á toma de decisións en canto á detección dunha determinada enfermidade. Consecuentemente, resulta interesante introducir os valores preditivos (positivo e negativo), que miden a eficiencia real dunha proba diagnóstica, proporcionando a probabilidade de padecer, ou non, unha determinada enfermidade, en función da prevalencia (probabilidade pre–proba) da doenza no total da poboación (ou nun rango específico de idade, no caso de restrinxir o noso estudo a un

subconxunto de interese da poboación). Deste modo, o valor preditivo positivo (PPV_+) defínese como a probabilidade de estar enfermo se o resultado da proba é positivo, é dicir:

$$\mathrm{PPV}_+ = \frac{\mathrm{VP}}{\mathrm{VP}+\mathrm{FP}}, \tag{3.2}$$

e o valor preditivo negativo (PPV_-) defínese como a probabilidade de estar san se o resultado da proba diagnóstica é negativo, isto é:

$$\mathrm{PPV}_- = \frac{\mathrm{VN}}{\mathrm{VN}+\mathrm{FN}}. \tag{3.3}$$

Segundo o exposto no artigo orixinal de Altman and Bland (1994), as expresións (3.2) e (3.3) poden reformularse en termos de sensibilidade (Rec), especificidade (Esp) e prevalencia (Pre), obtendo que:

$$\mathrm{PPV}_+ = \frac{\mathrm{Rec}\cdot\mathrm{Pre}}{\mathrm{Rec}\cdot\mathrm{Pre}+(1-\mathrm{Esp})\cdot(1-\mathrm{Pre})}, \tag{3.4}$$

$$\mathrm{PPV}_- = \frac{\mathrm{Esp}\cdot(1-\mathrm{Pre})}{(1-\mathrm{Rec})\cdot\mathrm{Pre}+\mathrm{Esp}\cdot(1-\mathrm{Pre})}, \tag{3.5}$$

posto que a prevalencia se determina numericamente como

$$\mathrm{Pre} = \frac{\mathrm{VP}+\mathrm{FN}}{\mathrm{VP}+\mathrm{FP}+\mathrm{FN}+\mathrm{VN}}.$$

Na nosa fonte de datos, o valor da prevalencia (mostral) é Pre= 0.086 (% da clase **FA**).

A diferenza de métricas propias dunha proba diagnóstica –como a sensibilidade, a especificidade, a medida $F1$, etc.– os valores preditivos (positivo e negativo) avalían o comportamento do modelo de diagnose nunha poboación específica, logo de determinar a proporción de suxeitos enfermos no total do censo. Consecuentemente, estes dous valores proporcionan probabilidades post–proba que serven para medir a relevancia da sensibilidade e da especificidade en certa poboación.

En liñas xerais, débese ter coidado coas medidas que usan como estimación da probabilidade de positivo a prevalencia mostral, porque poden non reflexar fielmente o que ocorre coa poboación de estudo, por exemplo, sobrerrepresentando a clase de interese. Unha posible solución é utilizar valores teóricos coñecidos. No noso caso, cunha sensibilidade e especificidade medias de 0.828 e 0.9892, respectivamente, e unha prevalencia teórica global[1] oscilando entre 0.01 e 0.02, obtemos que:

$$(\mathrm{PPV}_+^{\mathrm{Global}}, \mathrm{PPV}_-^{\mathrm{Global}}) = (0.4364, 0.9982) \quad \text{e} \quad (\mathrm{PPV}_+^{\mathrm{Global}}, \mathrm{PPV}_-^{\mathrm{Global}}) = (0.6101, 0.9965). \tag{3.6}$$

Como resultado, o $\mathrm{PPV}_+^{\mathrm{Global}}$ oscila entre 0.4364 e 0.6101, con valores moi afastados en relación aos esperados para o $\mathrm{PPV}_-^{\mathrm{Global}}$, que se move entre 0.9965 e 0.9982. Frecuentemente, a elevada proporción de suxeitos sans incrementa o número de falsos positivos, o que xustifica que o PPV_+ tenda a ser baixo se a prevalencia da enfermidade na poboación de estudo tamén o é. Polo contrario, un maior número de suxeitos enfermos aumenta a cantidade de diagnósticos falsos negativos, o que explica o descenso do PPV_- cando a prevalencia da enfermidade aumenta. As igualdades da expresión (3.6) confirman como, efectivamente, a prevalencia inflúe nos valores preditivos como se acaba de razoar.

Finalmente, debemos comentar que en poboacións de alto risco o resultado cambia notablemente. Recordando que a prevalencia teórica asociada a suxeitos maiores de 80 anos ascende a 0.15, e restrinxindo a poboación de estudo a esa franxa de idade, obtemos que os novos valores preditivos son

$$\mathrm{PPV}_+^{\mathrm{Maiores}} = 0.9312 \quad \text{e} \quad \mathrm{PPV}_-^{\mathrm{Maiores}} = 0.9703,$$

aumentando notablemente a probabilidade de padecer a enfermidade logo dun resultado positivo na proba. Nesta situación, menos dun 7 % dos resultados positivos e dun 3 % dos negativos son erróneos.

[1] En relación á notación adoptada, os valores preditivos correspondentes ao total do censo veñen acompañados do superíndice "Global" e os referidos á poboación maior de 85 incorporan o superíndice "Maiores".

Capítulo 4

Conclusións e traballo futuro

O presente traballo afronta a clasificación automática de rexistros de ECG, priorizando o cribado de FA, mediante a incorporación de novos modelos de aprendizaxe profunda, como son as redes neuronais Bi-LSTM e os mecanismos de atención. En definitiva, o estudo do rendemento do clasificador proposto permítenos afirmar que a aproximación presentada acada moi bos resultados, competitivos cos existentes ata o de agora na literatura. Sen embargo, e como ocorre na maioría de problemas de índole clínica, trátase dunha cuestión de difícil solución porque os sinais de ECG conteñen información descoñecida, complicada de modelar. Non obstante, a axeitada combinación de estatística e intelixencia artificial, xunto co inestimable traballo realizado por *Construe*, permiten que esta proposta conforme unha solución aceptable, tanto en termos de erro como no que afecta á interpretabilidade da solución.

Neste último aspecto, e baixo a certeza de que existen factores descoñecidos e difíciles de modelar que inflúen no etiquetado, debemos asumir que a imperfección dos modelos de clasificación é un factor practicamente incorrixible e que, dada unha solución aceptable en termos de erro, tamén é moi importante que sexa auto–explicable no sentido de que non usemos unha caixa negra para obter un bo modelo, principal problema das redes neuronais e, en xeral, da aprendizaxe profunda. Por este motivo, ademais de perseguir bos resultados numéricos, búscase acadar unha alta interpretabilidade da solución, razón pola cal se incorporan as redes de atención, principal aportación realizada.

4.1. Conclusións experimentais

A análise dos resultados permite afirmar que o cribado de FA é un problema complexo, confundíndose con outras patoloxías ou incluso con rexistros procedentes de pacientes sans. En canto a isto, determínase que gran parte da dificultade radica na identificación correcta doutros ritmos anómalos e en conseguir modelar os seus patróns máis característicos, probablemente polo feito de que a categoría **O** engloba arritmias diversas, como a taquicardia, a bradicardia, distintos tipos de extrasístoles (en función do número de latexos involucrados), os bloqueos de rama, etc. Sen embargo, a metodoloxía aplicada proporciona resultados competitivos cos presentados ata agora na literatura[1] e é unha primeira aproximación que dota de interpretabilidade a elección da etiqueta final.

Os resultados oficiais das puntuacións $F1$ dos primeiros posicionados no *Physionet*/CinC *Challenge* 2017 (Clifford et al., 2017), calculados sobre o conxunto de adestramento (8528 rexistros públicos que usamos na aprendizaxe) son, en maior ou menor medida, comparables cos obtidos por nós sobre o mesmo conxunto. Sen embargo, estas cantidades tenden a estar sobredimensionadas (adestramos e validamos sobre os mesmos datos) e é preferible utilizar un conxunto distinto, nunca visto, sobre o

[1] Como recurso de comparación do modelo proposto con outros clasificadores, centrarémonos en aqueles con rendemento superior. No Apéndice A recóllense os resultados oficiais máis relevantes do *Physionet*/CinC *Challenge* 2017, así como certas puntuacións $F1$ de CV para o modelo exposto en Teijeiro et al. (2018b), posteriores á fase final oficial.

que aplicar as métricas de desempeño. Como xa sabemos, no reto tiñan acceso a un conxunto de test oculto, que é para o que se facilitan as medidas $F1$ do Apéndice A. Sen embargo, para lograr unha comparación rigorosa, debemos calcular as medidas sobre os mesmos *folds* de validación, pero isto non está ao noso alcance para os modelos participantes, agás para o traballo de Teijeiro et al. (2018b).

Os resultados obtidos indican que o clasificador final é un bo candidato para utilizarse en dispositivos portátiles de diagnóstico e monotorización cardíaca, constituíndo un modelo competitivo, con medidas $F1$ de validación superiores ás dos gañadores do *Physionet*/CinC *Challenge* 2017. Neste aspecto, a idea é mellorar a detección prematura de arritmias, o que permitiría un tratamento precoz e un mellor seguimento, reducindo o risco de padecer graves complicacións e o custe sanitario asociado. Sen embargo, existen problemas de interpretabilidade que amosan a necesidade dun laborioso traballo futuro para conseguir un clasificador bo en termos estatísticos e de plena interpretabilidade da solución.

Ata o de agora, este obxectivo acadouse parcialmente, descubrindo unha serie de problemas repetitivos nalgúns rexistros con factores rítmicos e/ou morfolóxicos comúns. Trátase de complicacións derivadas dunha falta de acordo entre os criterios dos anotadores, de erros na detección de latexos e na mala identificación de ritmos de *Construe*, que se traducen nunha difícil barreira que impide a mellora dos resultados, independentemente da alta capacidade preditiva e de xeneralización do modelo. Ademais, estas complicacións non só son responsables da mala clasificación dalgunhas referencias senón que ocasionan que os modelos aprendan erroneamente certos patróns e fundamentan conxecturas equívocas sobre o esperable en rexistros relativos a pacientes sans/enfermos.

A continuación, recóllese a lista cos obstáculos descubertos e coa súa casuística. Convén destacar que todos son problemas de interpretabilidade de difícil solución e cuxo erro asociado debemos asumir.

I. Bradicardia: discordancia cardiolóxica cos limiares.

   Esta arritmia pode ser un grave problema de saúde se o corazón non bombea suficiente sangue ou non enriquece axeitadamente o corpo de osíxeno pero, nalgunhas persoas –especialmente en adultos novos, sans, e atletas en activo–, non se considera unha enfermidade porque non causa síntomas nin complicacións. En moitos casos, a variabilidade presente á hora de decidir se é unha razón para determinar se un paciente padece unha doenza cardíaca conduce a desacordo, o que impide unha boa clasificación en presenza de ritmos bradicárdicos predominantes.

II. Extrasístole: discrepancia cardiolóxica cando a ocorrencia é esporádica e problemas de *Construe*.

   Este incidente pode evidenciar a presenza dunha doenza cardíaca ou pode presentarse frecuentemente en corazóns sans. É esencial analizar a existencia doutras condicións patolóxicas que poidan acompañalo. Afirmamos que existen opinións contraditorias en canto aos criterios de anotación, o que ocasiona erros na clasificación de rexistros maioritariamente normais cunha extrasístole illada. Por outra banda, moitas das extrasístoles identificadas en rexistros normais débense a unha detección prematura do latexo, forzándose cambios na frecuencia cardíaca ficticios.

III. Asistolia: confusión de *Construe* coa perda de latexos.

   Unha asistolia é un descubrimento electrocardiográfico grave, incompatible con **N** e **FA**. Sen embargo, identifícase en exceso: detéctase incorrectamente ao comezo de bastantes rexistros, por fallos no detector de latexos en sinal ruidosa, e é común en rexistros de baixa amplitude, causados por un mal manexo do dispositivo de medición. Isto ocasiona que o modelo non aprenda o carácter patolóxico das asistolias e que, do mesmo xeito que a bradicardia e as extrasístoles, se asocie tamén a pacientes sans, o que prexudica gravemente a idea de modelo interpretable.

IV. Bloqueo de ritmo: discrepancia cardiolóxica cando a ocorrencia é esporádica e confusións de *Construe* coa perda de latexos e os cambios de ritmo.

   Un bloqueo de ritmo é un evento patolóxico que ocasiona que ao corazón lle custe bombear a sangue eficazmente, pero tamén pode presentarse nalgún paciente san. Existen anotacións

contraditorias en relación aos bloqueos pero, o principal problema, son erros de *Construe*. Cando se perde un latexo, detéctase un bloqueo seguido dunha asistolia, o que prexudica o significado de ambos ritmos. Ademais, o patrón de bloqueo de *Construe* actual só incorpora a lonxitude do segmento RR do latexo no que se identifica o bloqueo, o que desacertadamente conduce a detectalos en exceso: boa parte dos bloqueos identifícanse nun cambio de ritmo normal a bradicárdico, e viceversa. En raras ocasións, tamén ocorre involucrando ritmos taquicárdicos.

V. Latexos ventriculares: problema de detección de *Construe*.

   Os latexos ventriculares son frecuentes tanto en pacientes sans como con cardiopatías (causando certas arritmias ventriculares malignas, eventos de morte súbita e síncopes) e diagnostícanse en base á análise dun ECG: móstrase un complexo QRS ancho, sen onda P precedente e seguido dunha pausa compensadora completa. Máis de 2-3 latexos ventriculares en 30 segundos é considerado patolóxico e non pode acompañar á etiqueta **N**. O problema é que *Construe* non percibe o carácter ventricular de moitos latexos e, sen esta información, clasifícanse rexistros **O** como **N**.

Resumidamente, os problemas anteriores son externos á aprendizaxe dos clasificadores –causados por falta de acordo entre os cardiólogos e equivocacións de *Construe*– e non deben considerarse fallos derivados do modelo proposto. En realidade, grazas a este traballo proporciónanse direccións de mellora na definición interna dos procesos involucrados en *Construe*, enfocadas á detección de latexos e a unha caracterización rítmica dos mesmos máis exhaustiva. En xeral, enténdese que un rexistro se corresponde co esperable dun paciente san se presenta un ritmo regular constante no tempo e a morfoloxía é non patolóxica e estable. Adicionalmente, a ausencia de acordo na etiquetaxe demostra que outorgar completa independencia aos cardiólogos provoca resultados inconsistentes.

4.2. Intentos de mellora

Nesta sección preséntanse varios intentos de mellora que, apoiándonos nos resultados de validación, non superaron o rendemento do clasificador proposto. Por esta razón, non se continuou indagando nesas direccións nin veñen acompañados dunha comparación/explicación exhaustiva. A maiores, todos supoñen un aumento inxustificado da complexidade do modelo: sen melloras na clasificación e que, eventualmente, ocasionan barreiras na interpretabilidade.

En definitiva, ningunha das vindeiras modificacións reporta progresos nos resultados e, por este motivo, non constitúe unha alternativa competitiva en relación ao modelo proposto.

i. Substituír as celdas LSTM por celdas *peephole* (Gers and Schmidhuber, 2000).

   A súa estrutura é similar pero engaden conexións *peephole*, o que dota á rede da capacidade de contar eventos e duracións distanciados no tempo: permiten que as celdas non só dependan do anterior estado oculto ($\vec{h}_{t-1}$) senón que tamén influía a anterior (sub)celda de memoria (C_{t-1}).

   É razoable pensar que axudan se as engadimos á capa de ritmos, permitindo medir eventos e duracións (como a proporción de latexos do rexistro interpretados como ritmo normal, a cantidade de extrasístoles dun fragmento, a duración dunha bradicardia ou taquicardia, etc.). Modificaríase a arquitectura da rede do Diagrama 2.5, substituíndo as celdas LSTM desa capa.

ii. Incorporación das morfoloxías dos latexos.

   Estas caracterizan a súa silueta, estrutura, etc., o que xustifica a súa integración nunha nova capa de atención, con ou sen celdas *peephole*. Sen embargo, logo de comprobar que non reporta melloras nos resultados, conclúese que o comportamento morfolóxico dun latexo se modela co resto de variables secuenciais. Engadir as morfoloxías implica manexar 28 variables máis.

   Trasladar a variable **profile** á capa de ritmos ou á de morfoloxías –como indicador da calidade do sinal–, para modelar cales son *acertados* e cales están afectados polo ruído, non produce

o resultado esperado: nin mellora a clasificación nin solventa problemas de interpretabilidade, senón que incrementa o número de rexistros mal clasificados, o que indica unha peor aprendizaxe.

iii. Aumento do número de vectores de atención.

A súa contía é clave na complexidade e interpretabilidade da rede, determinando a cantidade de parámetros a aprender e a necesidade de engadir unha penalización por dependencia lineal das filas da matriz de pesos (para evitar que repitan os mesmos patróns).

Os mellores resultados reportados proceden de considerar un vector de atención por capa. Analizándoo detidamente, considerar máis prexudica rexistros sinxelos –predominantes en **N**–, aínda beneficiando aos máis complexos –máis frecuentes en **O** e **FA**– porque completar as filas da matriz de peso require esaxerar información, forzando á rede a atender a eventos irrelevantes.

iv. Reexecución de *Construe*.

Eliminación automática dos primeiros latexos identificados erroneamente como asistolias e reamplificación do sinal cando sexa moi baixa (ocupe o cuantil 0.01), redefiníndoa como o promedio da amplitude orixinal e a mediana da amplitude prereescalado.

Empeoran os resultados, obtendo por unha banda rexistros máis curtos e por outra, rexistros reamplificados tal que a detección de latexos continúa fallando.

v. Colapso de morfoloxías ventriculares.

Redúcense as morfoloxías nunha variable binaria 0-1 (**morph_colapse**) que modela o carácter ventricular. Determínase que as morfoloxías 4, 10 e 22 son ventriculares, recodificándose a 1, e enténdese que ás restantes se lles atribúe o valor 0. Realízanse distintas variacións na estrutura das redes: engadir **morph_colapse** á capa de ritmos e manter intacta a xeral; engadir **morph_colapse** á capa xeral e manter intacta a de ritmos; e crear unha nova capa de atención destinada á *análise ventricular*, coas variables **morph_colapse**, **QRSd⩾ 110** e **RR**.

En ningún caso se resolve o problema da detección de latexos ventriculares. Ademais, é un erro compartido co clasificador global por ser orixinal de *Construe*.

4.3. Traballo futuro

Moitas adaptacións e arquitecturas de rede neuronal, intentos de mellora na interpretabilidade e exploracións do espazo de variables, deixáronse para un futuro próximo debido á falta de tempo. Non obstante, este traballo contén unha análise profunda do modelo presentado, discutindo a súa elección en detrimento doutras alternativas similares e analizando as súas debilidades. Por esta razón, un labor recomendable sería continuar certas liñas de traballo que quedaron pendentes para o futuro.

Con posterioridade, sería beneficioso indagar na busca de novas características (globais e secuenciais) que axuden a mellorar a discriminación de rexistros etiquetados como doutras patoloxías e realizar un proceso de selección de variables máis rigoroso e pormenorizado, co fin de eliminar todas aquelas que non teñan unha aportación real ao rendemento do clasificador, para obter un modelo máis sinxelo e interpretable, en especial, tamén en termos da clasificación global. Igualmente, sería interesante validar o modelo proposto noutras bases de datos: ao incrementar o volume e diversidade dos datos, poderíamos mellorar a capacidade de xeralizar, atallando o problema máis temido da aprendizaxe profunda –e en particular, das redes neuronais artificiais–, o sobreaxuste.

Para terminar, co fin de reafirmar a mellora que supón a nosa aportación e salientar a súa aplicabilidade na detección precoz de arritmias cardíacas a nivel ambulatorio, con especial atención no cribado de FA, sería transcendental realizar un ensaio clínico. Para que os resultados froito deste estudo conformasen un conxunto de evidencias concluíntes na posterior tarefa de validación, reducindo a poboación

de interese aos maiores de 65 anos e cunha prevalencia estimada do 5 %, necesitaríamos 2766 pacientes para garantir unha sensibilidade de 0.9 e unha especificidade de 0.95, ao 5 % de significación, segundo as relacións expostas en Buderer (1996). Para o cálculo deste tamaño mostral, debemos determinar:

- O nivel de significación, $\alpha = 0.05$, que fixa a precisión clínica aceptable.
- A prevalencia da enfermidade na poboación de interese, Pre= 0.05 (suxeitos maiores de 65 anos).
- A sensibilidade esperada no novo test diagnóstico, Rec=0.9.
- A especificidade esperada no novo test diagnóstico, Esp= 0.95.

Posteriormente, debemos calcular o número de suxeitos enfermos, VP+FN,

$$\mathrm{VP} + \mathrm{FN} = z^2_{\alpha/2} \frac{\mathrm{Rec}(1-\mathrm{Rec})}{\alpha^2} = 138.2925,$$

onde $z_{\alpha/2}$ é o cuantil que deixa cola $\alpha/2$ á esquerda dunha distribución N(0,1), que no noso caso se aproxima por 1.96, e o tamaño mostral necesario para a sensibilidade:

$$N_1 = \left\lceil \frac{\mathrm{VP} + \mathrm{FN}}{\mathrm{Pre}} \right\rceil = 2766,$$

onde $\lceil x \rceil$ denota á parte enteira por exceso dun número real x.
A continuación, debemos calcular o número de suxeitos sans, FP+VN,

$$\mathrm{FP} + \mathrm{VN} = z^2_{\alpha/2} \frac{\mathrm{Esp}(1-\mathrm{Esp})}{\alpha^2} = 72.98772$$

e o tamaño mostral necesario para a especificidade:

$$N_2 = \left\lceil \frac{\mathrm{FP} + \mathrm{VN}}{1 - \mathrm{Pre}} \right\rceil = 77.$$

En último lugar, debemos seleccionar o tamaño mostral (N) do futuro ensaio clínico. Isto é, debemos coñecer o número de suxeitos precisos para estimar a sensibilidade e a especificidade conxuntamente, nunha poboación cunha prevalencia teórica do 5 %, común ao nivel de significación establecido. En concreto, esta cantidade calcúlase como o máximo entre N_1 e N_2.
No noso caso, segundo as cantidades prefixadas, $N = \mathrm{máx}\{N_1, N_2\} = N_1 = 2766$ suxeitos.
Finalmente, destacar que reducindo o nivel de significación do 5 % ao 1 %, o tamaño mostral demandado, segundo as igualdades expostas en Buderer (1996), ascende drasticamente, obtendo que

$$N_1 = 119429, \quad N_2 = 3318 \ \text{ e } \ N = 119429 \ \text{ suxeitos.}$$

Apéndice A

Physionet/CinC *Challenge* 2017

Neste apéndice coméntanse os resultados oficiais relativos ao *Physionet*/CinC *Challenge* 2017 (Clifford et al., 2017) e inclúense os valores $F1$ obtidos ao aplicar 8-*fold* CV ao modelo proposto no estudo de investigación de Teijeiro et al. (2018b), que constitúen unha fonte de comparación clave.

En primeiro lugar, a puntuación ordenada das 75 propostas independentes presentadas ao *Physionet*/CinC *Challenge* 2017 está dispoñible no enderezo

https://physionet.org/content/challenge-2017/1.0.0/results.csv,

que recolle a ordenación dos participantes en función da medida $F1$ global sobre o conxunto de test (redondeada ás dúas cifras decimais), así como o valor acadado para o conxunto de adestramento e o nome completo dos autores. Para coñecer os valores $F1$ de **N**, **FA** e **O** debemos visitar o enderezo

https://physionet.org/content/challenge-2017/1.0.0/results_all_F1_scores_for_each_classification_type.csv,

que ademais de incluír a información anterior, contén os valores **F1N**, **F1FA**, **F1O** –relativos á medida $F1$ individual para as clases **N**, **FA** e **O**– sobre os conxuntos de test e adestramento.

Na Táboa A.1 recóllese unha listaxe comparativa na que aparecen as mellores participacións, segundo o rendemento acadado no conxunto de test, incluíndo a medida $F1$ colectiva e os seus valores individuais sobre cada unha das clases do conxunto $\{\mathbf{N}, \mathbf{FA}, \mathbf{O}\}$. Neste caso, podemos observar que a proposta de Teijeiro et al. (2018b) se posiciona entre as catro con mellor desempeño, sendo a que acada un valor $F1$ máis elevado para a clasificación de fibrilación auricular.

Posición	**F1N**	**F1FA**	**F1O**	**F1score**	Autores
1	0.9039	0.8547	0.7366	0.83	Tomás Teijeiro; Constantino A. García; Paulo Félix; Daniel Castro
1	0.9158	0.8225	0.7498	0.83	Shreyasi Datta; Chetanya Puri; Ayan Mukherjee; Rohan Banerjee; Anirban Dutta Choudhury; Arijit Ukil; Soma Bandyopadhyay; Rituraj Singh
1	0.9087	0.8351	0.7341	0.83	Morteza Zabihi; Ali Bahrami Rad
1	0.9117	0.8128	0.7505	0.83	Shenda Hong; Yuxi Zhou; Qingyun Wang; Meng Wu; Junyuan Shang

Sigue na páxina seguinte.

Posición	**F1N**	**F1FA**	**F1O**	**F1score**	Autores
5	0.9107	0.8250	0.7299	0.82	Mohammed Baydoun; Lise Safatly; Hassan Ghaziri; Ali El-Hajj
5	0.9090	0.8221	0.7319	0.82	Martin Zihlmann; Michael Tschannen; Dmytro Perekrestenko
5	0.9112	0.8191	0.7328	0.82	Guangyu Bin; Minggang Shao; Jiao Huang; Guanghong Bin
5	0.9031	0.8203	0.7310	0.82	Zhaohan Xiong; Dr Jichao Zhao

Táboa A.1: Ordenación das primeiras 8 puntuacións dos gañadores do *Physionet*/CinC *Challenge* 2017, relativas á fase final. Os valores **F1N**, **F1FA**, **F1O** e **F1score** obtéñense coa versión V3 do reetiquetado, sobre o conxunto oculto de test (3658 sinais de ECG).

O conxunto de test, como comentamos ao longo do traballo, non está dispoñible para validar novos modelos externos aos presentados ao desafío e, polo tanto, non foi posible o seu uso na validación do rendemento do noso clasificador. Sen embargo, as anteriores medidas constitúen unha fonte de información que permite comparar os nosos resultados de validación cruzada coas mellores aportacións realizadas ao reto. De feito, en canto á métrica $F1$ de calidade da clasificación, promediando os resultados de validación cruzada (8-*fold* CV), obtemos que ascende a 0.9079 para o conxunto $\{\mathbf{N}, \mathbf{FA}, \mathbf{O}\}$, aumentando aínda máis para **N**, con un valor de 0.9714. Para **FA**, recordemos que chegaba a 0.8534, sendo o valor máis pequeno, seguido de 0.8989 para **O**.

Nunha etapa posterior á resolución do reto, realizouse unha fase de seguimento na que os participantes dispoñían da oportunidade de mellorar o seu desempeño anterior, perfeccionando as súas propostas. Deste xeito, e como se expón en Teijeiro et al. (2018b), a medida $F1$ para este equipo –sobre o conxunto de test– mellorou ata 0.85, logo dunha significativa simplificación do modelo (co obxectivo de reducir o sobreaxuste e incrementar a xeneralización), acadando a puntuación máis alta reportada ata agora para o conxunto de test. Os resultados de 8-*fold* CV para este último modelo –sobre o conxunto de adestramento e calculados nas mesmas condicións que os expostos no Cadro 3.2 para a nosa proposta– recóllense no Cadro A.1.

	Fold number								
Method	0	1	2	3	4	5	6	7	Mean (SD)
XGBoost	0.889	0.874	0.862	0.866	0.871	0.905	0.908	0.867	0.880 (0.018)
LSTMs	0.866	0.868	0.849	0.862	0.848	0.870	0.886	0.862	0.864 (0.012)
LDA-stacker	0.904	0.883	0.872	0.887	0.872	0.901	0.905	0.886	0.889 (0.013)

Cadro A.1: Resultados do modelo presentado en Teijeiro et al. (2018b) e dos dous clasificadores de nivel inferior: LDA-*stacker* é o modelo apilado, XGBoost é o clasificador global e *LSTMs* é o clasificador secuencial. Medidas $F1$, media e desviación típica nos 8 *folds* de CV estratificada.

Unha análise comparativa dos Cadros 3.2 e A.1 reflexa o deterioro do XGBoost, por mor da selección de variables, coa idea de diminuír dependencias/colinealidades e dotar ás redes de máis poder na clasificación, reducindo o solapamento de variables globais e secuenciais de natureza similar, cando é preferible medir esta latexo a latexo, e non globalmente. Así, para o clasificador global, redúcese a medida $F1$ media nos 8 *folds* de 0.880 a 0.8666, duplicándose a desviación típica. Contrariamente, e a pesar da redución do espazo de variables secuenciais en 7 unidades, o comportamento das

redes mellora, ascendendo de 0.862 a 0.8997, cunha desviación típica moi similar. En particular, observamos como o promedio dos valores $F1$ de validación esta próximo a 0.90, mellorando individual e exitosamente ao modelo *ensemble* de Teijeiro et al. (2018b). Por último, o rendemento do modelo apilado increméntase –a causa de modificar o clasificador *ensemble* e da incorporación das redes de atención–, ascendendo de 0.889 a 0.9079, cunha leve diminución da desviación típica, de 0.013 a 0.0044.

En vista dos resultados estatísticos presentados, e apoiándonos na medida $F1$ de CV sobre $\{\mathbf{N}, \mathbf{FA}, \mathbf{O}\}$, concluímos que o modelo proposto mellora ao exposto en Teijeiro et al. (2018b) en canto a súa capacidade de xeralización e, como xa comentamos, tamén en termos da interpretabilidade dos resultados.

Apéndice B

Variables de entrada dos clasificadores

Como se comentou ao longo do traballo, para a extracción de características globais e secuenciais, comúns as empregadas en Teijeiro et al. (2018b), usouse o *framework Construe* (Teijeiro and Félix, 2018). A continuación, imos presentar unha descrición máis detallada, dividíndoas segundo sexan variables de entrada do clasificador global ou secuencial, con especial atención a ritmos e morfoloxías.

B.1. Variables globais

Construe extrae un conxunto de 42 características globais, descritas no Cadro B.1[1] coa idea de entender que información aportan ao modelo –no caso de ser incorporadas– e sinálase cales se usaron finalmente na clasificación global e a causa de exclusión das que foron descartadas. A súa función é modelar medidas resumo de todo o rexistro, que permitan detectar comportamentos patolóxicos prolongados no tempo ou tal que sexa a súa elevada frecuencia a causa da anomalía, como ocorre coa notable irregularidade do ritmo cardíaco propia da fibrilación auricular.

Variable global	Breve descrición	Decisión final codificada
tSR	Proporción da duración do rexistro interpretado como ritmo regular (ritmo normal, taquicardia ou bradicardia)	1
t1b	Milisegundos dende o comezo ata o primeiro latexo interpretado	0
tOR	Número de milisegundos interpretados como ritmo non regular	1
longTch	Período temporal máis longo con frecuencia cardíaca superior a 100 latexos por minuto	2
RR	Media do intervalo RR en ritmos regulares	2
RRd_std	Desviación típica da variación instantánea do intervalo RR	2
RRd	DAM* do intervalo RR en ritmos regulares	2
MRRd	Máxima variación absoluta do intervalo RR en ritmos regulares	2

Sigue na páxina seguinte.

[1] A descrición das variables globais presentada é unha tradución ao galego da exposta en Teijeiro et al. (2018b).

Variable global	Breve descrición	Decisión final codificada
RR_MIrr	Máxima medida de irregularidade do intervalo RR	2
RR_Irr	Mediana da irregularidade do intervalo RR	2
PNN10	Medida PNN10 global	1
PNN50	Medida PNN50 global	1
PNN100	Medida PNN100 global	1
o_PNN50	Medida PNN50 de ritmos non regulares	1
mRR	Mínimo intervalo RR de ritmos regulares	1
o_mRR	Mínimo intervalo RR de ritmos non regulares	1
n_nP	Proporción de latexos con onda P detectada en ritmos regulares	0
n_aT	Mediana da amplitude das ondas T en ritmos regulares	0
n_PR	Duración media do intervalo PR en ritmos regulares	2
Psmooth	Mediana do cociente entre a desviación estándar e a media do sinal derivada das ondas P	0
Pdistd	DAM* da medida dada polo método de delineación da onda P	0
MPdist	Máximo da medida dada polo método de delineación da onda P	1
prof	Perfil do sinal completo	2
pw_profd	DAM* de **pw_prof** (variable secuencial)	2
xcorr	Mediana da máxima correlación cruzada entre todos os pares de complexos QRS interpretados en ritmos regulares	2
o_xcorr	Mediana da máxima correlación cruzada entre todos os pares de complexos QRS interpretados en ritmos non regulares	2
PRd	DAM* global das duracións do intervalo PR	2
QT	Mediana de QTc, a medida QT corrixida	1
TP	Mediana da frecuencia predominante nos intervalos TP	2
TPfreq	Mediana da entropía de frecuencia nos intervalos TP	2
pw_prof	Perfil do sinal na área da onda P	2

Sigue na páxina seguinte.

Variable global	Breve descrición	Decisión final codificada
nT	Proporción de complexos QRS con ondas T detectadas	1
n_Txcorr	Mediana da correlación cruzada máxima entre todos os pares de ondas T en ritmos regulares	0
n_Pxcorr	Mediana da correlación cruzada máxima entre todos os pares de ondas P en ritmos regulares	0
baseline	Perfil da liña de base en ritmos regulares	2
o_baseline	Perfil da liña de base en ritmos non regulares	2
wQRS	Proporción de complexos QRS amplos (duración superior a 110 ms)	1
wQRS_xc	Mediana da máxima correlación cruzada entre todos os pares de complexos QRS amplos	2
wQRS_prof	Mediana do sinal nos 300 ms previos a cada complexo QRS amplo	1
w_PR	Proporción de latexos con intervalo PR longo (máis de 210 ms)	1
x_xc	Mediana da máxima correlación cruzada entre todos os pares de latexos ectópicos	2
x_rrel	Mediana do cociente entre a lonxitude dos intervalos RR anterior e seguinte a cada latexo ectópico	2

*Observación: DAM significa desviación absoluta media.

Cadro B.1: Conxunto de variables globais e resultados da selección de variables. En maxenta (0), as eliminadas no Paso 1; en marrón (1), as eliminadas no Paso 2; e en oliva (2), as restantes variables.

Algunhas variables globais do Cadro B.1 requiren unha explicación máis pormenorizada:

i. O termo *profile* –utilizado en **prof**, **pw_prof**, **wQRS_prof**, **baseline** e **o_baseline**– refírese á suma dos valores absolutos da desviación do sinal e resulta un excelente indicador da súa calidade.

ii. O método de delineación da onda P –mencionado para as variables **Pdist** e **MPdist**– é un procedemento incluído en *Construe* que mide cando se parece a onda P observada a unha estándar.

iii. As medidas **pNNx**, con $x \in \{10, 50, 100\}$ e **o_pNN50**, relativas á variabilidade da frecuencia cardíaca, explícanse no artigo de investigación de Mietus et al. (2002) e son cruciais na detección de certas arritmias, entre as que se atopa a FA. O reconto NNx defínese como o promedio de veces por hora en que os cambios de intervalos consecutivos de ritmo normal (de aí o termo NN) superan os x milisegundos. Tendo isto en valor, defínese a medida **pNNx** como

$$\textbf{pNNx} = \frac{\text{reconto NNx}}{\text{reconto total NN}},\ x \in \{10, 50, 100\},$$

que representa a porcentaxe de intervalos NN consecutivos que difiren en máis de x milisegundos. Do mesmo xeito, defínese a medida **o_pNN50** como

$$\textbf{o\_pNN50} = \frac{\text{reconto NN50 baixo ritmo non regular}}{\text{reconto total NN baixo ritmo non regular}}.$$

Unha serie de probas ensaio–erro conducen á eliminación destas 4 variables globais no Paso 2, concluíndo que a variabilidade da frecuencia cardíaca é máis conveniente modelala latexo a latexo, na clasificación secuencial (analizando individualmente o comportamento dos intervalos RR), e non globalmente (con medidas resumo dos mesmos).

B.2. Variables secuenciais

Construe extrae un conxunto de 38 variables secuenciais, que ascenden a 43 pola definición de variables auxiliares, que modelan limiares fisiolóxicos importantes. Estas ocúpanse de medir características básicas do latexo, que permiten detectar desviacións do esperable en pacientes sans ou indicios de certas arritmias, como a fibrilación auricular. Ademais, identifícanse 12 ritmos e 28 morfoloxías. No Cadro B.2 detállanse estas $12 + 28 + 43 = 83$ características destinadas a modelar a información secuencial dos rexistros de ECG, xunto cos resultados codificados do proceso de selección de variables.

Variable temporal	Índice	Breve descrición	Decisión final codificada
Rh	0:11	Ritmo cardíaco. Considéranse 12, dende o 0 ata o 11, codificados como 12 características binarias 0-1	2
Morph	-	Morfoloxías do complexo QRS. Considéranse 28, codificadas como 28 características binarias 0-1	1
w1detected	12	Binaria 0-1: Indicadora de detectar máis dunha onda no complexo QRS (detéctanse 2-3 ondas)	2
w2detected	13	Binaria 0-1: Indicadora de detectar máis de dúas ondas no complexo QRS (detéctanse 3 ondas)	2
Pwdetected	14	Binaria 0-1: Indicadora de detectar a onda P	2
Twdetected	15	Binaria 0-1: Indicadora de detectar a onda T	2
TPdetected	16	Binaria 0-1: Indicadora de detectar a onda TP, que non se observa nun ECG normal	2
RR⩾ 1200	17	Binaria 0-1 auxiliar: Indicadora de que a lonxitude do segmento RR sexa superior a 1200	2
RR⩽ 600	18	Binaria 0-1 auxiliar: Indicadora de que a lonxitude do segmento RR sexa inferior a 600	2
QRSd⩾ 110	19	Binaria 0-1 auxiliar: Indicadora de que a duración do complexo QRS sexa superior a 110	0
PR⩾ 210	20	Binaria 0-1 auxiliar: Indicadora de que a lonxitude do segmento PR sexa superior a 210	2
RRda⩽ 55	21	Binaria 0-1 auxiliar: Indicadora de que a diferenza do RR entre latexos consecutivos sexa inferior a 50	2
Rpk	22	Tempo do latexo (detección do complexo QRS)	2
RR	23	Lonxitude do intervalo RR ou primeiro latexo	2
RRdb	24	Diferenza entre a lonxitude do RR actual e anterior	2

Sigue na páxina seguinte.

Variable temporal	Índice	Breve descrición	Decisión final codificada
RRda	25	Diferenza entre a lonxitude do RR seguinte e actual	2
RRIrr	26	Medida de estimación da entropía	2
w0a	27	Amplitude da primeira onda do complexo QRS	2
w0d	28	Duración da primeira onda do complexo QRS	2
w0p	29	Distancia entre a detección da primeira onda do complexo QRS ao instante onde acada o seu pico	2
w1a	30	De detectarse, amplitude da segunda onda do QRS	2
w1d	31	De detectarse, duración da segunda onda do QRS	0
w1p	32	De detectarse, distancia entre a detección da segunda onda do QRS ao instante onde acada o seu pico	2
w2a	33	De detectarse, amplitude da terceira onda do QRS	2
w2d	34	De detectarse, duración da terceira onda do QRS	0
w2p	35	De detectarse, distancia entre a detección da terceira onda do QRS ao instante onde acada o seu pico	2
Axis	36	Eixo do complexo QRS	2
QRSd	37	Duración do complexo QRS	2
QRSa	38	Amplitude do complexo QRS	2
pw_prof	39	Perfil de área da onda P	2
PR	40	Lonxitude do segmento PR	2
Pwd	41	Duración da onda P	2
Pwa	42	Amplitude da onda P	0
Pwdist	43	Distancia da onda P actual á seguinte	0
Twd	44	Duración da onda T	0
Twa	45	Amplitude da onda T	2
QT	46	Lonxitude do intervalo QT corrixido, QTc	2
STdev	47	Desviación do segmento ST	2

Sigue na páxina seguinte.

Variable temporal	Índice	Breve descrición	Decisión final codificada
TPa	48	Amplitude da onda TP	2
TPf	49	Frecuencia da onda TP	0
TPfa	50	Amplitude espectral da onda TP	2
atrial_entr	51	Medida de entropía da frecuencia en actividade auricular	2
TPdur	52	Duración da onda TP	0
profile	53	Perfil do latexo (indicador da súa calidade)	2
baseline	54	Lonxitude da liña de base do ECG (liña isoeléctrica)	2

Cadro B.2: Conxunto de variables secuenciais e resultados da selección de variables. En maxenta (0), as eliminadas en base á idea de atención; en marrón (1), as morfoloxías; e en oliva (2), as restantes.

Moitas variables secuenciais son unha desagregación das súas homólogas globais, como ocorre con **RR**, **RRIrr**, **n_PR**, **QT**, **TP**, **Pdistd**, **prof**, **pw_prof** e **TPfreq**. Tamén se consideran medidas morfolóxicas como a duración, amplitude e punto de inflexión de cada onda nun latexo. Así mesmo, **Rh** modela o patrón rítmico e **Morph** modela o patrón morfolóxico (dentro do marco abdutivo de interpretación de *Construe*). Ambas terán subseccións destinadas a súa explicación detallada.

Por outra banda, as redes neuronais son modelos moi sensibles á escala das variables, esixindo a súa homoxeneización. Isto débese a súa formulación matemática que, recordemos, se fundamenta en combinacións lineais que logo se transforman mediante funcións de activación diferenciables. En consecuencia, é imprescindible a tarefa de preprocesado das variables de entrada.
No noso caso, realizouse coas seguintes funcións da libraría *sklearn* de `Python` (2022):

- *StandardScaler*: Estandariza as variables continuas, restando a media e escalando, para obter variables centradas na orixe e con varianza unitaria. Permite unificar magnitudes e escalas e, polo tanto, realiza o necesario proceso de homoxeneización.
- *OneHotEncoder*: Codifica as variables categóricas como unha matriz numérica con entradas binarias que indican a que categorías se corresponden. Cada categoría modélase como un vector da mesma dimensión que o número de elementos diferentes que existan (o vocabulario), con todas as entradas iguais a 0 agás a que a represente, que estará definida a 1.

Para manexar unha base de datos axeitada, tamén se substitúen os datos faltantes. Ademais, e como xa comentamos, aplícanse funcións lóxicas relativas a limiares fisiolóxicos, obtendo novas variables.

B.2.1. Ritmos

Construe realiza unha interpretación a nivel de ritmo de cada latexo, asignando unha etiqueta que pode ser de ritmo normal ou dalgún dos 11 ritmos patolóxicos definidos –como se precisa no Cadro B.3– e que se asocian coas diferentes arritmias e eventos anómalos máis comúns. Ademais, a súa importancia é clave na interpretación dos resultados das redes de atención porque unha das capas que conforman a arquitectura do clasificador secuencial está exclusivamente destinada á análise rítmica.

Codificación *Construe*	Rítmo cardíaco	Breve descrición
0	Normal/Sinusal	Ritmo cardíaco normal
1	Bradicardia	Pulsacións inferiores a 60 latexos/minuto
2	Taquicardia	Pulsacións superiores a 100 latexos/minuto
3	Extrasístole	Latexo adiantado respecto á frecuencia cardíaca normal
4	Doblete ou parella	Dúas contraccións ventriculares prematuras
5	Bigeminismo	Contracción ventricular prematura cada 2 latexos
6	Trigeminismo	Contracción ventricular prematura cada 3 latexos
7	Bloqueo de ritmo	Retraso do latexo ocasionado por demora ou fallos na condución
8	Asistolia	Período de ausencia de latexo: Carencia de actividade eléctrica no miocardio. Principal causa de paro cardíaco
9	Flúter ventricular	Frecuencia cardíaca superior a 200 latexos/minuto, presentando un patrón continuo sinusal, sen que se visualicen correctamente os complexos QRS e as ondas T
10	Fibrilación auricular	Ritmo irregular, con tempo entre latexos desigual
11	Descoñecido	Ritmo non identificado como familiar

Cadro B.3: Explicación dos 12 ritmos definidos, recollidos na variable secuencial **Rh**.

B.2.2. Morfoloxías

Construe identifica 28 morfoloxías distintas do complexo QRS, caracterizadas en función da duración, tamaño e amplitude das ondas Q, R e S que o conforman, aplicando unha terminoloxía intuitiva apoiada no uso de letras maiúsculas e minúsculas. Por exemplo:

- qRs: onda Q inicial pequena (non patolóxica), seguida de onda R alta e onda S pequena.
- Rs: onda R alta seguida de onda S pequena.
- Qr: onda Q profunda seguida de onda R pequena.

Cando algunha das ondas falta, prodúcese un incidente anómalo que debemos ter en conta na clasificación e que está modelado nas variables secuenciais **w1detected** e **w2detected**.

O total de morfoloxías detectadas vén determinado polo conxunto {Q, QR, QRs, QRS, QS, Qr, QrS, Qs, R, RR, RS, RSR, Rr, RrS, Rs, RsR, qR, qRs, qS, qr, r, rR, rS, rSr, rr, rs, rsR, rsr} e, para facilitar a notación, asóciase nunha relación directa biunívoca cos índices $\{0, 1, ..., 27\}$.

Debemos destacar que, ademais das morfoloxías en si, defínense variables secuenciais cuxa natureza é totalmente morfolóxica, como as destinadas a medir a amplitude, duración e detección das ondas, intervalos e segmentos dun latexo (representados na Figura 1.3). Isto xustifica porque, en definitiva, as 28 morfoloxías se descartan na construción do clasificador secuencial.

Apéndice C

Adestramento e optimización

Neste apéndice explícanse os labores de adestramento e optimización dos parámetros e hiperparámetros relativos aos modelos de clasificación presentados e aplicados no presente traballo. Por unha parte, os parámetros (estruturais) son os pesos que debemos aprender no adestramento e que, posteriormente, serán os responsables da clasificación de elementos nunca vistos. En consecuencia, obtéñense como resultado da aprendizaxe dos modelos e nunca se definen manualmente. No noso caso, atopámonos con modelos paramétricos porque a cantidade de parámetros a calcular –tanto no XGBoost como nas redes neuronais e nas SVM– son valores fixos. Por outra parte, os hiperparámetros (parámetros de axuste ou *tuning parameters*) son cantidades establecidas *a priori*, e que caracterizan as configuracións utilizadas no adestramento, impoñendo restricións estruturais ao modelo como, por exemplo, o número de parámetros. Habitualmente, obtéñense mediante procesos de optimización que involucran á busca con validación cruzada. Por esta razón, a optimización realízase entre un conxunto de valores posibles espazados nunha grella, ao que nos referimos como enfoque *greedy search*. En particular, cando a busca se realiza entre todos os valores, fálase dun enfoque *exhaustive greedy search*, e cando só se avalían algunhas combinacións aleatorias, fálase de *randomized greedy search*. No noso caso, seguiuse un enfoque *exhaustive greedy search*, máis custoso computacionalmente pero que garante que se comproben todos os posibles cruces de hiperparámetros definidos segundo a nosa grella.

Do mesmo xeito que na inferencia estatística clásica, un modelo demasiado complexo padece sobreaxuste e, polo contrario, un demasiado simple sofre infraaxuste (*underfitting*). Por esta razón, é clave no rendemento do modelo unha axeitada selección dos seus hiperparámetros.

En primeiro lugar, comezaremos co clasificador global e, seguidamente, abordaremos o modelo secuencial. Por último, explicarase o adestramento e optimización do clasificador *ensemble*. No noso caso, o elevado número de hiperparámetros do XGBoost e, máis aínda, das redes neuronais, dará lugar a problemas de optimización complicados, con solucións aproximadas e computacionalmente custosas.

C.1. Clasificador global

O XGBoost incorpora hiperparámetros propios dos modelos de aprendizaxe baseados en árbores de decisión e outros máis específicos, que permiten controlar o sobreaxuste. Moitos deles resultan cruciais para o seu rendemento, como a máxima profundidade de cada árbore; a taxa de aprendizaxe, que escala os pesos das mesmas para reducir a súa influencia na aprendizaxe e deixar marxe de mellora para que futuras árbores poidan aportar; o factor que controla o número de ramificacións de cada árbore (penaliza complexidade); a proporción de características usadas e escollidas ao azar en cada árbore, imitando aos bosques aleatorios (*random forest*), introducidos por primeira vez no artigo de Breiman (2001); o peso mínimo das árbores e o mínimo número de elementos da mostra para construír cada unha, etc. No noso caso, a optimización dos hiperparámetros do modelo global realizouse seguindo

un enfoque *exhaustive greedy search*, con 60 rondas de *boosting*, equivalentes a un adestramento con 60 iteracións, obtendo os resultados recollidos no Cadro C.1.

Hiperparámetro	Breve descrición	Valor
max_depth	Profundidade máxima de cada árbore	5
η	Taxa de aprendizaxe	0.2
γ	Factor que controla o número de nodos de cada árbore	1.0
colsample_bytree	Proporción de variables usadas ao azar en cada árbore	0.9
min_child_weigth	Peso mínimo de cada árbore	20
subsample	Proporción mostral utilizada para construír cada árbore	0.8

Cadro C.1: Hiperparámetros do clasificador global.

Os pesos dos datos mostrais determináronse tal que os relativos ás clases máis representadas contaran cunha penalización maior con respecto aos relativos ás clases con menor representación. Concretamente, seguiuse unha estratexia habitual de ponderación tal que se calcula o peso do rexistro $i-$ésimo como a diferenza entre a unidade e a proporción mostral de datos coa mesma etiqueta. Recordando que n denotaba ao tamaño mostral, defínese c^i como a clase do rexistro $i-$ésimo ($c^i \in \{\mathbf{N}, \mathbf{FA}, \mathbf{O}, \smile\}$) e $\mathbb{I}(c^j = c^i)$ como a variable aleatoria indicadora de que a clase do rexistro $j-$ésimo coincide coa clase c^i. Como resultado, o peso do rexistro $i-$ésimo vén dado por

$$w_i = 1 - \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}(c^j = c^i),\ i = 1, ..., n.$$

Neste escenario, resulta máis doado traballar directamente cos pesos dos datos en función da súa clase, e non co rexistro en cuestión. Deste xeito, os pesos utilizados no adestramento do XGBoost son[1]

$$w_N = 1 - p_N = 0.35982, \quad w_{FA} = 1 - p_{FA} = 0.91403,$$
$$w_O = 1 - p_O = 0.77023 \ \text{ e } \ w_\smile = 1 - p_\smile = 0.96376.$$

Como consecuencia, os rexistros ruidosos e os etiquetados como **FA** terán máis peso á hora de adestrar ao clasificador global (a súa contribución á función de perda será maior e, polo tanto, resultará máis importante lograr a súa boa clasificación), a causa da súa baixa proporción mostral. Este feito é de gran relevancia se recordamos os criterios externos á aprendizaxe (referentes ás clases $\smile$ e **FA**) formulados na Sección 2.3, adicada ao clasificador *ensemble*. Por outra banda, o peso asociado aos rexistros etiquetados como **O** duplica ao da clase **N**, favorecendo a correcta discriminación dos primeiros.

Finalmente, o método de resolución aplicado (que neste contexto se chama *booster*) foi o *gradient tree boosting*, fixado coa opción *booster='gbtree'*, e coincide co exposto en teoría. Como comentamos na

[1]Non reporta ningún problema que os pesos usados no adestramento do XGBoost, que determinan a contribución das observacións (segundo a clase) na construción do modelo global, non sumen 1. A normalización realizase internamente no algoritmo de *boosting* implementado en `Python` (2022). Esixindo que o vector $(w'_N, w'_{FA}, w'_O, w'_\smile)$ sume 1, obtemos:

$$w'_N = 0.11963, \quad w'_{FA} = 0.30388, \quad w'_O = 0.25607, \quad w'_\smile = 0.32042.$$

Sección 2.1, a ecuación utilizada no proceso de minimización polo cal se engaden aditivamente clasificadores débiles ao modelo vén dada pola expresión (2.7). A función de perda é a *softprob* multiclase, cun comportamento similar á función *softmax* –que se introduce na ecuación (C.1) para a optimización de redes neuronais–, diferenciándose na saída. Coa opción *softmax* obtense a clase máis probable para cada observación, e coa opción *softprob*, un vector de probabilidades repartidas entre as clases. En consecuencia, pódese interpretar como a distribución de probabilidade sobre K posibles saídas mutuamente excluíntes: a cada observación asígnaselle unha probabilidade distinta de pertencer a cada unha das clases, determinando a etiqueta final en función do máximo destas cantidades.

C.2. Clasificador secuencial

As arquitecturas baseadas en redes neuronais artificiais e, en particular, as redes neuronais recorrentes e as redes LSTM, incorporan unha serie de hiperparámetros cruciais para o seu rendemento. Seguidamente, defínense os conceptos máis relevantes do seu adestramento e optimización:

- O tamaño do lote (*batch size*) é o número de elementos da mostra que conforman as particións que se propagarán en conxunto a través da rede, nunha etapa. A medida que aumenta, máis espazo de memoria se require. Tipicamente, as redes adéstranse máis rápido se o seu valor é máis pequeno. Isto é porque actualizamos os pesos despois de cada iteración.
- O número de etapas (*number of epochs*) é o máximo de iteracións realizadas no adestramento, utilizando todos os datos, divididos en lotes.
- A taxa de aprendizaxe (*learning rate*) é o escalar que multiplica ao gradiente no método de descenso do gradiente estocástico e modela a velocidade da aprendizaxe. É clave nos resultados do algoritmo de *backpropagation*. Un valor moi pequeno tradúcese nunha aprendizaxe demasiado lenta. A maior valor, máis rápida será a aprendizaxe, pero pode empeorar o adestramento, impedir a converxencia a cero do erro e favorecer o sobreaxuste do modelo.
- O valor de parada (*early stopping value*) é o escalar que establece baixo que limiar se forza unha parada prematura do adestramento, como causa do incremento do erro de validación.
- O valor de abandono (*dropout value*) é o escalar que indica que proporción de neuronas serán omitidas aleatoriamente nunha etapa, co fin de reducir o sobreaxuste.
- Os valores de decaemento dos pesos (*weight decay values*) son cantidades que modelan a penalización de pesos grandes, engadindo termos propios da regularización *ridge* (penalización en norma $\mathcal{L}^2$) e *lasso* (penalización en norma $\mathcal{L}^1$) á función de perda da rede neuronal.

Para obter unha boa clasificación e evitar sobreaxustar os datos de adestramento, debemos usar mecanismos de parada, escoller unha taxa de aprendizaxe axeitada, incluír penalizacións aos pesos, apagar neuronas aleatoriamente, etc. Co fin de establecer os valores dos hiperparámetros, podemos fixalos de antemán ou empregar métodos de optimización. Neste caso, os hiperparámetros optimizáronse segundo un enfoque *exhausitive greedy search*, excepto o tamaño de lote e o número de etapas que, xunto coa arquitectura da rede neuronal, se fixaron de antemán a 64 e 20, respectivamente. Durante a tarefa de adestramento, aplicouse 8-*fold* CV, obtendo os resultados almacenados no Cadro C.2, distinguindo entre os hiperparámetros procedentes da capa de ritmos e os relativos á capa xeral.

Hiperparámetro	Breve descrición	Valor
lr [RedLR]	Taxa de aprendizaxe	0.00255

Sigue na páxina seguinte.

Hiperparámetro	Breve descrición	Valor
lr_patience [RedLR]	Número de iteracións sen melloras ata reducir lr	3
lr_factor [RedLR]	Factor que, multiplicativamente, reduce lr: lr_new=lr·lr_factor	$\sqrt{0.5}$
lr_min [RedLR]	Límite inferior de lr (valor mínimo)	10^{-7}
min_delta [EarS]	Limiar para medir melloras significativas	10^{-6}
patience [EarS]	Nº de iteracións sen melloras ata forzar a parada	15
lstm_drp	Proporción de neuronas a esquecer no adestramento da entrada dunha celda LSTM*	Ritmos: 0.3 Xeral: 0.25
lstm_recurrent_drp	Proporción de neuronas a esquecer no adestramento do estado recurrente dunha celda LSTM*	Ritmos: 0.275 Xeral: 0.25
lstm_l2_alpha	Regularización aplicada aos pesos dunha celda LSTM*	Ritmos: 0.0025 Xeral: $5 \cdot 10^{-5}$
l2_alpha	Regularización aplicada aos pesos da saída ANN	$3.7625 \cdot 10^{-3}$
ann_drp	Proporción de neuronas a esquecer no adestramento na saída dunha celda ANN	0.25
*Observación:	O valor do hiperparámetro depende da capa de atención	Ritmos Xeral

Cadro C.2: Hiperparámetros do clasificador secuencial.

A maiores dos hiperparámetros do Cadro C.2, o número de unidades ocultas das dúas capas Bi-LSTM fixouse a 128, aumentando ata 256 na capa ANN final. As funcións de activación de cada camiño, posteriores ás redes Bi-LSTM, correspóndense coa tanxente hiperbólica e as funcións de activacións que proporcionan os vectores de atención e a clasificación final escolléronse nos tres casos iguais á función *softmax*. Neste sentido, debemos destacar que a función de activación *softmax*, ou función exponencial normalizada, é a xeralización da regresión loxística a datos continuos e máis de dúas clases. En concreto, dado un vector de valores reais, devolve as probabilidades de cada compoñente sobre o total das posibles, tomando valores no intervalo $(0, 1]$ e sumando 1.
Formalmente, dado un vector $\overrightarrow{x} = (x_1, ..., x_K) \in \mathbb{R}^K$,

$$\begin{aligned} f(\overrightarrow{x}) =& (f_1(\overrightarrow{x}), ..., f_K(\overrightarrow{x})) \in (0, 1]^K, \text{ con} \\ & f_k(\overrightarrow{x}) = \frac{e^{x_k}}{\sum_{j=1}^{K} e^{x_j}} \in (0, 1], \ k = 1, ..., K \text{ compoñentes (no noso caso, clases).} \end{aligned} \tag{C.1}$$

Por outra banda, o optimizador usado na aprendizaxe implementa o algoritmo de Adam –variante do descenso do gradiente estocástico–, que se basea na estimación de momentos de primeira e segunda orde. A función de perda é a entropía cruzada categórica (*categorical cross entropy*), que calcula a perda de entropía cruzada entre as etiquetas e as clases preditas polo modelo.

Ademais, engádase un mecanismo de parada prematura (*EarlyStopping*, EarS), que usa o valor da función de perda no conxunto de validación para regular o adestramento: interrompe a aprendizaxe cando a función de perda deixa de mellorar. Posteriormente, incorpórase un mecanismo de mellora tal que se reduce a taxa de aprendizaxe cando a función de perda sobre o conxunto de validación deixa de

diminuír (*ReduceLROnPlateau*, RedLR), co idea de ralentizar a velocidade do adestramento para evitar ignorar mínimos locais próximos ao valor da función de perda na correspondente iteración. Por último, unha vez finalizado o proceso iterativo, compróbanse e compáranse automaticamente os resultados intermedios relativos aos pesos obtidos en cada iteración (*ModelCheckpoint*), en canto ao valor da función de perda computada sobre o conxunto de validación, quedándonos cos pesos correspondentes ao modelo con mellor desempeño (coa perda sobre o conxunto de validación menor).

C.3. Clasificador *ensemble*

Na configuración da máquina de soporte vectorial utilizada como metaclasificador na etapa de *ensemble*, elixiuse o núcleo lineal e a optimización do hiperparámetro de regularización realizouse seguindo un enfoque *exhausitive greedy search*, concluíndo que $C = 0.7$. Para esta tarefa, aplicouse validación cruzada de 8 pregaduras, considerando como entrada as probabilidades do XGBoost e das redes de atención. Debemos acentuar que a magnitude da regularización é inversamente proporcional a C e sempre debe ser estritamente positiva, sendo esta a regularización cadrática (regularización *ridge*). Primeiramente, a busca dun valor axeitado para C iniciouse no conxunto

$$\{0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 5, 7.5, 10\},$$

onde a mellor cantidade resultou ser $C = 0.75$ e, posteriormente, refinouse co conxunto

$$\{0.6, 0.65, 0.7, 0.75, 0.8, 0.85\}.$$

Finalmente, quedámonos co valor $C = 0.7$.

Para terminar, comentar que se probaron outros tipos de núcleo –como o sigmoide, o polinómico ou o exponencial, para distintos graos do polinomio– pero os resultados de validación cruzada acadados eran peores que os obtidos con núcleo lineal. A continuación, defínense as funcións núcleo mencionadas, aplicadas a dous vectores X e Y (de dimensión arbitraria D pero común):

1. Núcleo lineal: a función defínese como o produto escalar $K(X, Y) = X^t Y$.
2. Núcleo sigmoide: $K(X, Y) = \tanh(1 + \gamma X^t Y)$, con γ un hiperparámetro adicional a determinar e que modela a importancia do produto escalar $X^t Y$ na aplicación da tanxente hiperbólica.
3. Núcleo polinómico: $K(X, Y) = (1 + \gamma X^t Y)^d$, con γ e $d \in \mathbb{N}^{\geqslant 1}$ dous hiperparámetros adicionais a determinar e que modelan, respectivamente, a importancia de $X^t Y$ e o grao do polinomio.
4. Núcleo exponencial: $K(X, Y) = \mathrm{e}^{-\gamma \|X-Y\|^2}$, con $\gamma > 0$ un hiperparámetro adicional a determinar e que modela a curtose da función exponencial.

No noso caso, X e Y son vectores 6−dimensionais, é dicir, $X, Y \in \mathbb{R}^6$, e almacenan as probabilidades do clasificador global e secuencial para as clases **N**, **FA** e **O**. Por último, debemos destacar que todos os núcleos presentados, agás o núcleo lineal, dan lugar a fronteiras de decisión non lineais, caracterizándose a súa non linealidade segundo os valores dos respectivos hiperparámetros adicionais engadidos.

Bibliografía

Altman, D. G. and Bland, J. M. (1994). *Statistics Notes. Diagnostic tests 2: Predictive values.* British Medical Journal, 309:102.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). *Neural machine translation by jointly learning to align and translate.* Presented paper at the International Conference on Learning Representations.

Bishop, C. (2006). *Pattern Recognition and Machine Learning.* Springer.

Breiman, L. (2001). *Random forests.* Machine Learning, 45 (1), p. 5-32.

Buderer, N. (1996). *Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity.* Academic Emergency Medicine, 3(9), p. 895-900.

Chen, T. and Guestrin, C. (2016). *XGBoost: A scalable tree boosting system.* Proceedings of the 22th International Conference on Knowledge Discovery and Data Mining, p. 785-794.

Clifford, G., Liu, C., Moody, B., Silva, I., Li, Q., Johnson, A., and Mark, R. (2017). *AF classification from a short single lead ECG recording. The PhysioNet Computing in Cardiology Challenge (CinC).* Computing in Cardiology, 44.

Cui, Z., Ke, R., Pu, Z., and Wang, Y. (2016). *Deep stacked bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction.* 6th International Workshop on Urban Computing 2017.

Dalal, S. and Vishwakarma, V. P. (2021). *Classification of ECG signals using multi-cumulants based evolutionary hybrid classifier.* Scientific Reports, 11 (1): 15092.

Dean, J. and Monga, R. (2015). *TensorFlow: Large-scale machine learning on heterogeneous distributed systems.* TensorFlow.org. Google Research.

Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine.* The Annals of Statistics, 49 (5), p. 1189-1232.

Gers, F. and Schmidhuber, J. (2000). *Recurrent nets that time and count.* Proceedings International Joint Conference on Neural Networks 2000.

Goldberger, J. J., Challapalli, S., Waligora, M., Kadish, A. H., Johnson, D. A., Ahmed, M. W., and Inbar, S. (2000). *Uncertainty principle of signal-averaged electrocardiography: Components of a new research resource for complex physiologic signals.* Circulation, 101, p. 2909-2915.

Hastie, T., Tibshirani, R., and Friedman, J. (2017). *The Elements of Statistical Learning. Data Mining, Inference and Prediction.* Springer.

Hochreiter, S. (1998). *The vanishing gradient problem during learning recurrent neural nets and problem solutions.* International Journal of Uncertainty Fuzziness and Knowledge-Based Systems, 6(2), p. 107-116.

Hochreiter, S. and Schmidhuber, J. (1997). *Long short-term memory.* Neural Computing, 9(8), p. 1735-1780.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning with Applications in* `R`. Springer.

Lin, Z., Feng, M., dos Santos, C., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. (2017). *A structured self-attentive sentence embedding.* 5th International Conference on Learning Representations.

Lip, G., Fauchier, L., Freedman, S., Gelder, I., Natale, A., Gianni, C., Nattel, S., Potpara, T., Rienstra, M., Tse, H., and Lane, D. (2016). *Atrial fibrillation.* Nature Reviews Disease Primers, 2, 16016.

López de Ullibarri, I. and Pita, S. (1999). *Medidas de concordancia: El índice Kappa.* Cadernos de Atención Primaria, 6, p. 169-171.

Mietus, J. E., Peng, C., Henry, I., Goldsmith, R. L., and Goldberger, A. L. (2002). *The pNNx files: Re-examining a widely used heart rate variability measure.* Heart, 88(4), p. 378-380.

Misra, P. and Yadav, A. S. (2020). *Improving the classification accuracy using recursive feature elimination with cross-validation.* International Journal on Emerging Technologies, 11(3), p. 659-665.

Quinlan, J. (1986). *Induction of decision trees.* Machine Learning, 1(1), p. 81-106.

`R Core Team` (2022). *A language and environment for statistical computing.* `R` Foundation for Statistical Computing.

Rosenblatt, F. (1961). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.* Spartan Books. Washington, DC.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). *Learning internal representations by error propagation.* Explorations in the Microstructure of Cognition, 1: Foundations, p. 318-362.

Rumelhart, E., Williams, R., and Hinton, G. (1968). *Learning representations by back-propagating errors.* Nature, 323(6088), p. 533-536.

Savelieva, I. and Camm, J. (2008). *Update on atrial fibrillation.* Clinical Cardiology, 31(2), p. 55-62.

Schuster, M. and Paliwal, K. (1997). *Bidirectional recurrent neural networks.* IEEE Transactions on Signal Processing, 45(11), p. 2673-2681.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley and Sons, New Jersey, 2th edition.

Teijeiro, T. and Félix, P. (2018). *On the adoption of abductive reasoning for time series interpretation.* Artificial Intelligence, 262, p. 163-188.

Teijeiro, T., Félix, P., Presedo, J., and Castro, D. (2018a). *Heartbeat classification using abstract features from the abductive interpretation of the ECG.* IEEE Journal of Biomedical and Health Informatics, 22(2) p. 409-420.

Teijeiro, T., García, C., Castro, D., and Félix, P. (2018b). *Abductive reasoning as a basis to reproduce expert criteria in ECG atrial fibrillation identification.* Physiological Measurement, 39(8), 084006.

`Python` (2022). *A Programming Language.* `Python` Software Foundation.

Wolpert, D. H. (1992). *Stacked generalization.* Neural Networks, 5(2), p. 241-259.