تشخیص بیماری آریتمی قلبی با استفاده از PCA و شبکه عصبی MLP

ایرج نارویی[۱]، بهزاد زمانی[۲]

Diagnosis Of Cardiac Arrhythmia Using Pca And MLP Neural Network

Iraj Naruei, Behzad Zamani

Email: irajnaruei@iauzah.ac.ir

چکیده

بیماری‌های قلبی یکی از شایع‌ترین انواع بیماری‌ها است که آمار بسیار بالایی از مرگ و میر را به خود اختصاص می‌دهد به طوریکه سالانه حدود ۵۰۰۰۰۰ نفر بر اثر این بیماری جان خود را از دست می دهند. آریتمی ها ضربان‌های غیر طبیعی هستند، که موجب می‌شوند قلب خیلی سریع (تاکی کارد) یا خیلی آهسته (برادی کارد) بزند و پمپاژ غیر مؤثر داشته باشد. الکتروکاردیوگرام (ECG) پروسه بدون دردی است که فعالیت الکتریکی قلب را ضبط می کند. تجزیه وتحلیل خودکار ECG برای تشخیص ودرمان بیماران بدحال حیاتی است. ما در این تحقیق با استفاده از تحلیل مولفه اساسی[۳] (PCA) ۱۰۰ مولفه اول را به عنوان ورودی به شبکه عصبی با دو لایه پنهان که هر لایه ۱۰۰ نرون دارد وارد کردیم که برای طبقه بندی دودویی به دقت ۹۳,۴۰٪ و برای طبقه بندی چندکلاسه به دقت ۷۸٪ دست یافتیم. این آزمایشات روی مجموعه داده های آریتمی پایگاه UCI انجام شده است.

کلمات کلیدی

بیماری قلبی، تحلیل مولفه اساسی، شبکه عصبی، ECG، PCA

۱. مقدمه

آریتمی ها خیلی شایع بوده و سالانه میلیون ها نفر را در جهان درگیر می‌کنند.آنها علت اصلی مرگ ناگهانی قلبی در ایالات متحده هستند و سالیانه موجب ۴۰۰۰۰۰ مرگ می‌شوند.[1] بنابراین آریتمی‌ها می‌توانند تهدید کننده حیات باشند، البته در صورتی‌که کاهش شدیدی در عملکرد پمپاژی قلب ایجاد نمایند. وقتی عملکرد پمپاژی قلب به مدت بیش از چند ثانیه به شدت کاهش یافت، گردش خون ضرورتا قطع می‌شود و آسیب به ارگانها (مثل مغز) در عرض چند دقیقه بوجود می‌آید. در زمینه تشخیص آریتمی کارهای زیاد انجام شده که آقای صمد[۴] و همکارانش در سال ۲۰۱۴ از سه مدل K-nn ، Decision Tree و Naive Bayes برای طبقه بندی استفاده کردند که نرخ طبقه بندی به ترتیب ۶۶,۹۶٪ ، ۵۹,۷۷٪ و ۴۵,۸۵٪ حاصل شد.[2] آقای جدهو[۵] و همکارانش در سال ۲۰۱۰ یک شبکه

---

[۱] دانشجوی کارشناسی ارشد دانشگاه آزاد اسلامی واحد زاهدان

[۲] دکترای هوش مصنوعی و عضو هیات علمی دانشگاه علم و صنعت ایران

[3] Principal Component analysis

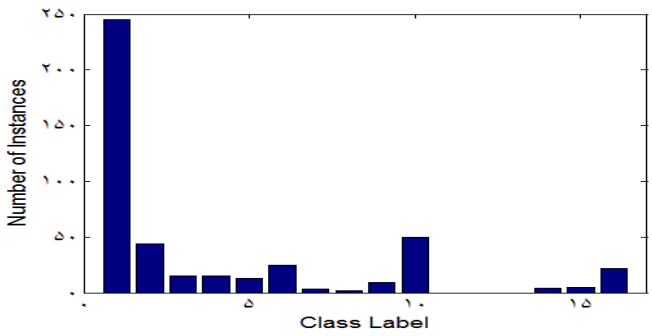
[4] Samad

[5] Jadhav

عصبی مدولار با تعداد لایه های پنهان مختلف از یک تا سه و با درصد آموزش مختلف در پارتیشن مجموعه داده ارائه دادند، مقادیر خصیصه های گم شده این مجموعه داده ها با مقادیر نزدیک ترین ستون کلاس مربوطه جایگزین شد. بهترین دقت طبقه بندی با ۲ لایه پنهان برای مجموعه داده پارتیشن ۵ برابر با ۸۲,۲۲٪ شد.[3] آقای جدهو و همکارانش در سال ۲۰۱۰ دوباره یک شبکه عصبی چند لایه با الگوریتم پس انتشار ارائه کردند که به دقت ۸۶,۶۷٪ رسیدند.[4] خانم مالای[6] و سامانتا در سال ۲۰۱۳ با استفاده از انتخاب ویژگی مبتنی بر همبستگی (CFS) ابعاد ویژگی را کاهش داده و برای طبقه بندی از شبکه عصبی پس انتشار تدریجی و الگوریتم LM استفاده کردند که برای هر طبقه بند به ترتیب به دقت ۸۶,۰۲٪ و ۸۷,۷۱٪ رسیدند.[5] آقای کوهلی[7] و ورما در سال ۲۰۱۱ از PCA برای استخراج ویژگی و از SVM به عنوان طبقه بند برای ۶ کلاس استفاده کردند که به دقتی بین ۸۲,۵۳٪ تا ۸۵,۷۱٪ دست یافتند. [6]

۲.مجموعه داده

مجموعه داده مورد استفاده در این تحقیق، از آرشیو مجموعه داده‌های یادگیری ماشین UCI گرفته شده است.[7] این مجموعه داده شامل ۴۵۲ نمونه و ۲۷۹ ویژگی است که برای سهولت استفاده از آن در تعیین وجود یا عدم وجود بیماری آریتمی و برای شناسایی نوع آریتمی به سه دسته کلی تقسیم شده است. در این مجموعه داده کلاس "۱" اشاره به ECG نرمال، کلاس‌های "۲-۱۵" اشاره به کلاس‌های مختلف آریتمی و کلاس "۱۶" به بقیه داده‌های طبقه‌بندی نشده اشاره دارد. توزیع نمونه ها در هر کلاس در شکل ۱ نشان داده شده است.



شکل ۱ : تعداد نمونه ها در هر کلاس

۳.پیش پردازش داده ها

در بسیاری از کاربردهای دنیای واقعی کاوش داده ها، حتی با وجود مقدار داده های حجیم و فضای ذخیره سازی مناسب، ممکن است در نمونه های موجود، مقادیری از داده ها از دست رفته(گمشده) باشند. اما مشکل از آنجا آغاز میشود که برای مجموعه داده های بزرگ نمی توان از مقادیر از دست رفته چشم پوشی کرد. یک راه حل برای جایگذینی خودکار مقادیر از دست رفته با مقادیر ثابت عبارت است از:

- جایگزینی تمام مقادیر از دست رفته با یک تک مقدار ثابت سراسری
- جایگزینی یک مقدار از دست رفته با متوسط مشخصه آن.
- جایگزینی یک مقدار از دست رفته با متوسط مشخصه آن برای یک گروه مشخص.

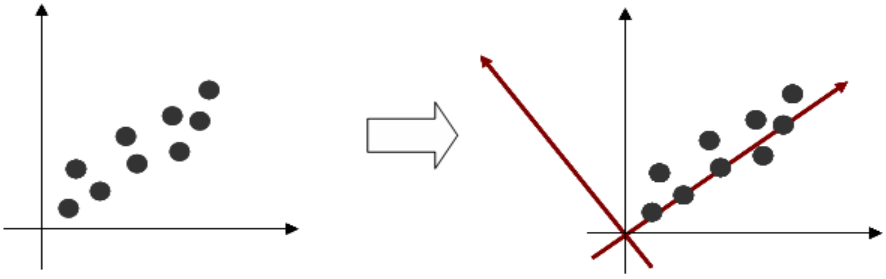
از آنجایکه در دیتاست ما ویژگی هایی با مقادیر مفقودی وجود دارد، و ما نمیخواهیم این اطلاعات را از دست دهیم، لذا مقادیر مفقودی با استفاده از روش میانگین مقادیر هر ویژگی مقدارگذاری می گردد.[8]

---

[6] Malay
[7] Kohli

۴. الگوریتم آنالیز اجزای اصلی (PCA)

تکنیک PCA بهترین روش برای کاهش ابعاد داده به صورت خطی می‌باشد. یعنی با حذف ضرایب کم‌اهمیت بدست آمده از این تبدیل، اطلاعات از دست رفته نسبت به روشهای دیگر کمتر است. البته کاربرد PCA محدود به کاهش ابعاد داده نمی‌شود و در زمینه‌های دیگری مانند شناسایی الگو و تشخیص چهره نیز مورد استفاده قرار می‌گیرد. در این روش محورهای مختصات جدیدی برای داده‌ها تعریف شده و داده‌ها براساس این محورهای مختصات جدید بیان می‌شوند. اولین محور باید در جهتی قرار گیرد که واریانس داده‌ها ماکسیمم شود (یعنی در جهتی که پراکندگی داده‌ها بیشتر است). دومین محور باید عمود بر محور اول به گونه‌ای قرار گیرد که واریانس داده‌ها ماکسیمم شود. به همین ترتیب محورهای بعدی عمود بر تمامی محورهای قبلی به گونه‌ای قرار می‌گیرند که داده‌ها در آن جهت دارای بیشترین پراکندگی باشند. در شکل۲ این مطلب برای داده‌های دو بعدی نشان داده شده است.



شکل۲: انتخاب محورهای جدید برای داده‌های دو بعدی

الگوریتم PCA به شرح زیر است:

M یک مجموعه داده t بعدی است. n محور اساسی G1,G2,…Gn بر هم عمود هستند. در حالت کلی، G1,G2,…,Gn می تواند به وسیله n بردار ویژه از ماتریس کوواریانس نمونه ها به دست آید.

$$C = \left(\frac{1}{L}\right)\sum_{k=1}^{L}(x_k - m)^T (x_k - m)$$

که $x_k \in M$، m میانگین نمونه ها و L تعداد نمونه ها است. با توجه به این مطلب:

$$UG_k = v_k G_k, k \in 1,...,n$$

که v_k، k تا بزرگترین مقدار ویژه U است. n جز اصلی یک بردار $x_k \in M$ به صورت زیر داده شده است:

$$q = [q_1, q_2, ..., q_n] = [G_1^T x, G_2^T x, ..., G_n^T] = G^T x$$

که q، n تا اجزای اصلی x می باشند.

۵. شبکه عصبی

این شبکه شامل سه لایه ورودی، مخفی و خروجی است که تعداد سلولهای هر لایه به روش سعی و خطا مشخص می گردد. سیگنال های ورودی به وسیله ضریب های بهنجار کننده به مقدار یک نرمالیزه شده و بعد از محاسبات، خروجی به مقدار واقعی برگردانده می شود. همچنین مقادیر اولیه وزن ها به صورت اتفاقی در نظر گرفته شده اند. این شبکه بر مبنای الگوریتم پس انتشار خطا اموزش می بیند. بدین ترتیب که خروجی های واقعی با خروجی های دلخواه مقایسه می شوند و وزن ها به وسیله الگوریتم پس انتشار، به صورت تحت نظارت تنظیم می گردند تا الگوی مناسب بوجود آید. برای الگوی ورودی p ام، مربع خطای خروجی برای تمامی سلول های لایه خروجی شبکه به صورت زیر در می آید:

$$E_p = \frac{1}{2}(d^p - y^p)^2 = \frac{1}{2}\sum_{j=1}^{s}(d_j^p - y_j^p)^2$$

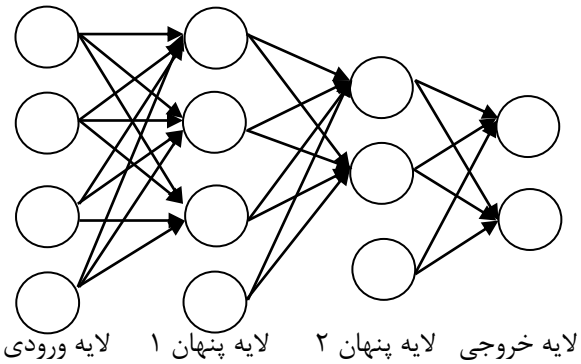
که در آن d_j^p خروجی دلخواه برای j امین سلول در لایه خروجی واقعی برای j امین سلول در لایه خروجی، s ابعاد بردار خروجی، y^p بردار خروجی واقعی و d^p بردار خروجی دلخواه هستند. مربع خطای کل E برای P الگو بصورت زیر در می آید:

$$E = \sum_{p=1}^{p} E_p = \frac{1}{2}\sum_{p=1}^{p}\sum_{j=1}^{s}(d_j^p - y_j^p)^2$$

وزن ها با هدف کاهش تابع هزینه E به مقدار مینیمم به روش گرادیان نزولی تنظیم می گردند. معادله به روز در آوردن وزن ها به صورت زیر است:

$$w_{ij}(t+1) = w_{ij}(t) + \eta\Delta w_{ij}(t) + \alpha\Delta w_{ij}(t-1)$$

که در آن $\Delta w_{ij}(t) = -(\frac{\partial E_p}{\partial w_{ij}(t)})$، η ضریب یادگیری، α ضریب لحظه ای $w_{ij}(t+1)$ وزن جدید و $w_{ij}(t)$ وزن قبلی می باشد. همچنین در این روش، وزن ها به طور مکرر برای تمامی الگوهای یادگیری به روز درآورده می شوند. روند یادگیری هنگامی متوقف می شود که مجموع کل خطا، E، برای p الگو از مقدار آستانه تعیین شده کمتر شود یا تعداد کل دوره تعلیم به پایان برسد. مدل شبکه عصبی ما شامل ۲ لایه پنهان با تعداد ۱۰۰ نرون برای هر لایه است، مدل شبکه عصبی ما بصورت شکل ۳ است.



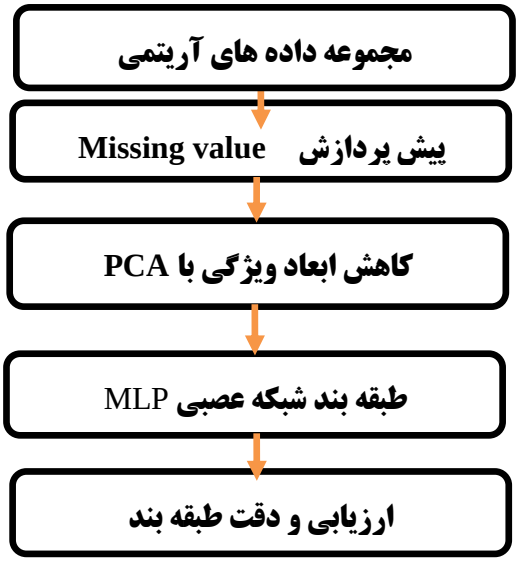
شکل ۳: معماری مدل شبکه عصبی MLP

ما یک شبکه عصبی چند لایه را برای هر دو طبقه بندی باینری و چند کلاسه بکار گرفتیم. خروجی هر نرون تابع سیگموئید به صورت زیر است.

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}, x = \begin{bmatrix} \overline{x_0} \\ x_1 \\ : \\ x_n \end{bmatrix}, \theta = \begin{bmatrix} \overline{\theta_0} \\ \theta_1 \\ : \\ \theta_n \end{bmatrix}$$

ما شبکه را با استفاده از پس انتشار بازگشتی و گرادیان تصادفی نزولی برای به حداقل رساندن تابع هزینه آموزش می دهیم. تنظیم پارامتر لامبدا برای جلوگیری از Overfitting با کاهش مقدار پارامتر در فرمول ۱ انجام می شود. برای تثبیت تمام پارامترهای دیگر، ما LAMDA را تغییر می دهیم و برابر با ۰،۰۱ تنظیم می کنیم.

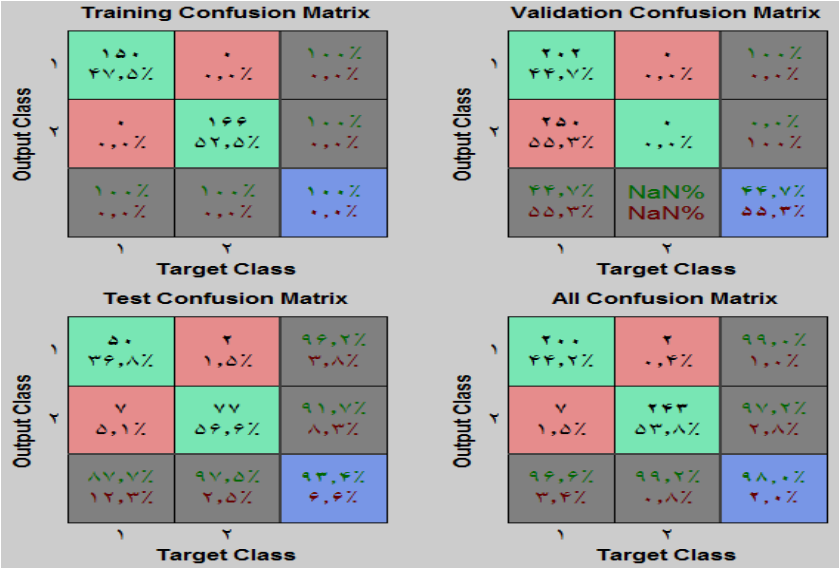
معماری مدل پیشنهادی ما در شکل ۴ ارائه شده است.



شکل ۴: مدل پیشنهادی

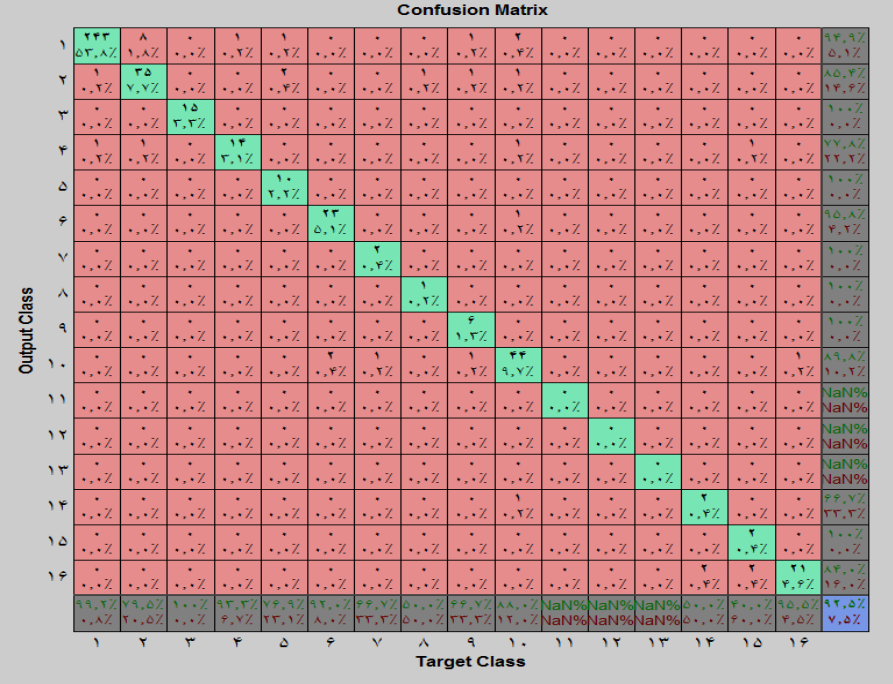
۶.نتایج

در مدل شبکه عصبی ۷۰٪ داده ها را برای آموزش و ۳۰٪ داده ها را برای تست در نظر گرفتیم . با استفاده از الگوریتم PCA صد مولفه اساسی را به عنوان ورودی به شبکه عصبی دادیم و از ۲ لایه پنهان با ۱۰۰ نرون برای شبکه عصبی در نظر گرفتیم که برای طبقه بندی دو کلاسه به دقت ۱۰۰٪ برای داده های آموزش و ۹۳,۴٪ برای داده های تست و دقت کلی ۹۸٪ دست یافتیم.



شکل۴: نتایج شبکه عصبی برای طبقه بندی دودویی

برای طبقه بندی چند کلاسه با ۷۰٪ داده ها برای آموزش و ۳۰٪ داده ها برای تست و دو لایه پنهان با ۱۰۰ نرون به دقت طبقه بندی ۱۰۰٪ برای داده های آموزش و ۷۸٪ برای داده های تست و دقت کلی ۹۲,۵٪ دست یافتیم.



شکل ۵: نتایج شبکه عصبی برای طبقه بندی چندکلاسه

در جدول ۱ نتایج ما با روش های دیگر مقایسه شده اند که روش پیشنهادی ما بالاترین دقت را دارد.

جدول ۱ : مقایسه روش ها

ردیف	روش	دقت طبقه بندی دودویی	دقت طبقه بندی چندکلاسه
۱	صمد و همکارانش سال ۲۰۱۴		
	K-nn	۶۶,۹۶	
	Decision Tree	۵۹,۷۷	
	Naïve Bayes	۴۵,۸۵	
۲	جدهو و همکارانش سال ۲۰۱۰		
	شبکه عصبی مدولار	۸۲,۲۲	
۳	جدهو و همکارانش سال ۲۰۱۰	۸۶,۶۷	
	شبکه عصبی چند لایه		
۴	مالای و سامانتا سال ۲۰۱۳		
	شبکه عصبی پس انتشار تدریجی	۸۶,۰۲	
	الگوریتم LM	۸۷,۷۱	
۵	کوهلی و ورما سال ۲۰۱۱		
	SVM-PCA برای ۶ کلاس		۸۲,۵۳ تا ۸۷,۷۱
۶	random forest	۹۰	۷۶
۷	**روش پیشنهادی ما**	**۹۳,٤**	**۷۸**

۷.نتیجه گیری

این نتایج مبین بهبودی قابل ملاحظه ای نسبت به کارهای اخیر در زمینه تفکیک آریتمی های قلبی است. توجه به این نکته ضروری است که اکثر مقالات یاد شده در بخش منابع بر روی همین پایگاه داده کار کرده اند و بیشتر آنها فقط به تفکیک دو کلاسه پرداخته اند، که با این وجود با نتایج این تحقیق قابل مقایسه نیستند. بکار گیری روش پیشنهادی می تواند کمک شایانی در زمینه تشخیص دقیق و صحیح امراض قلبی به پزشکان نماید.

۸.منابع و مراجع

[1] Sesselberg.HW, et al., 2007,Ventricular arrhythmia storms in post infarction patients with implantable defibrillators for primary prevention indications: A MADIT-II sub study,Heart Rhythm, Elsevier, Vol.4,1395–1402.

[2] Saleha Samad, et al.,2014, Classification of Arrhythmia, International Journal of Electrical Energy, Vol. 2, No. 1.

[3] Shivajirao M. Jadhav, et al., 2010a, ECG Arrhythmia Classification using Modular Neural Network Model, IEEE EMBS Conference on Biomedical Engineering & Sciences.

[4] Shivajirao M. Jadhav, et al., 2010b, Artificial Neural Network Based Cardiac Arrhythmia Classification Using ECG Signal Data, International Conference on Electronics and Information Engineering.

[5] Malay Mitra and R. K. Samanta, 2013, Cardiac Arrhythmia Classification Using Neural Networks with Selected Features, International Conference on Computational Intelligence, 76 – 84.

[6] Narendra Kohli and Nishchal K. Verma,2011, Arrhythmia classification using SVM with selected features, International Journal of Engineering, Science and Technology Vol. 3, No. 8, 2011, pp. 122-131

[7] Bache, K. and Lichman M., 2013, UCI machine learning repository,. URL http://archive.ics.uci.edu/ml.

[8] Jose M. Jerez, et al.,2010, Missing data imputation using statistical and machine learning methods in a real breast cancer pro