ORIGINAL ARTICLE

Iranian Quarterly Journal of Breast Disease 2018; 10(4):47.

Using Bio-geographical Algorithm in Optimizing Neural Network for the Diagnosis of Breast Cancer

Ahmadi Toussi C: Dept. of Medical Engineering, Faculty of electrical and computer science, Hakim Sabzevari University, Sabzevar, Iran
Ghayoumi-Zadeh H: Dept. of Medical Engineering, Faculty of electrical and computer science, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran.
Haddadnia J: Dept. of Medical Engineering, Faculty of electrical and computer science, Hakim Sabzevari University, Sabzevar, Iran

Corresponding Author: Cyrus Ahmadi Toussi**,** cyrus.ahmady@gmail.com

Abstract

Introduction: Breast cancer is the most common cancer in women. Accurate classification of breast cancer has a key role in medical diagnosis. Hence, researchers seek optimized methods to improve tumor diagnosis.

Methods: The current study presents bio-geographical based optimization neural network for classifying data as benign and malignant using principal component analysis in preprocessing stage and updating weights concurrently. The presented algorithm was assessed using the data from Wisconsin databank.

Results: Classification accuracy in a normal state, that is, without applying principal component analysis and an optimization algorithm, and applying only neural network at a ratio of %70 to %30 from training and testing set is %97.2. Accuracy reaches %98.5 after applying principal component analysis and decreasing features from nine to eight. Finally, using bio-geographical based optimization algorithm with a 10-fold cross validation, accuracy reaches %100, which is significantly more successful than other similar studies.

Conclusion: Applying this algorithm can optimize the performance of the neural network. The optimal performance of this method is revealed by comparing the proposed method with the non-optimized method and the approach which used only PCA and neural network method. The results suggest that the method presented in this paper had a high accuracy in classifying breast cancer data and can be used for its diagnosis.

Keywords: Bio-geographical Based Optimization, Principal Component Analysis, Multi-layer Perceptron, Back Propagation Method.

مقاله پژوهشی

فصلنامهٔ بیماری‌های پستان ایران، سال دهم، شماره چهارم، زمستان ۱۳۹۶؛ (۵۸-۴۸)

تاریخ ارسال: ۹۶/۸/۱۴ | تاریخ پذیرش: ۹۶/۱۱/۱

استفاده از الگوریتم جغرافیای زیستی در بهینه سازی شبکه عصبی جهت تشخیص سرطان پستان

سیروس احمدی طوسی[*]: گروه مهندسی پزشکی، دانشکده برق و علوم کامپیوتر، دانشگاه حکیم سبزواری، سبزوار، ایران
حسین قیومی‌زاده: گروه مهندسی پزشکی، دانشکده برق و علوم کامپیوتر، دانشگاه ولیعصر(ع) رفسنجان
جواد حدادنیا: گروه مهندسی پزشکی، دانشکده برق و علوم کامپیوتر، دانشگاه حکیم سبزواری، سبزوار، ایران

چکیده

مقدمه: در حال حاضر، سرطان پستان از شایع‌ترین بیماری‌های زنان است. دسته‌بندی دقیق تومور سرطان پستان نقش کلیدی را در امر تشخیص پزشکی ایفا می‌کند. متخصصین به دنبال روش‌های بهینه جهت بهبود تشخیص این تومور می‌باشند.
روش بررسی: در این مطالعه شبکه عصبی مبتنی بر جغرافیای زیستی ارایه گردیده که با استفاده از آنالیز اجزای اصلی در مرحله آماده‌سازی و بروز رسانی همزمان وزن‌ها موفق به دسته‌بندی داده‌ها به عنوان خوش‌خیم یا بدخیم می‌گردد. جهت ارزیابی الگوریتم ارایه شده از داده‌های بانک اطلاعاتی ویسکانسین استفاده شده است.
یافته‌ها: دقت تفکیک در حالت عادی یعنی حالتی که از آنالیز اجزای اصلی و الگوریتم بهینه سازی استفاده نشده و تنها شبکه عصبی با نسبت ۷۰-۳۰ داده‌های آموزش به تست مورد استفاده قرار گیرد، ۹۷/۲٪ است. با بکارگیری آنالیز اجزای اصلی و کاهش ۹ ویژگی به ۸ ویژگی دقت به ۹۸/۵ می‌رسد. نهایتا با استفاده از الگوریتم بهینه سازی جغرافیای زیستی همراه با اعتبار سنجی ضربدری ۱۰ گانه دقت به ۱۰۰٪ رسیده که به میزان قابل توجهی از نتایج بدست آمده از مطالعات دیگر موفق‌تر است.
نتیجه‌گیری: استفاده از این الگوریتم می‌تواند عملکرد شبکه عصبی را بهبود دهد. مقایسه روش ارایه شده با حالت بهینه نشده و در حالتی که فقط از PCA و شبکه عصبی استفاده شده است، عملکرد بهینه این روش را نشان داد. نتایج حاکی از آن است که مدل ارایه شده در این مقاله دقت بسیار بالایی در تفکیک داده‌های سرطان پستان دارا می‌باشد و می توان از آن جهت تشخیص نهایی این سرطان استفاده نمود.

واژه‌های کلیدی: بهینه سازی جغرافیای زیستی، آنالیز اجزای مستقل، شبکه عصبی پرسپترون چند لایه، الگوریتم پس انتشار.

[*] نشانی نویسنده مسئول: : خراسان رضوی، سبزوار، دانشگاه حکیم سبزواری، دانشکده فنی و مهندسی، گروه مهندسی پزشکی، سیروس احمدی طوسی.
نشانی الکترونیک: cyrus.ahmady@gmail.com

مقدمه

سرطان پستان از شایع‌ترین سرطان‌ها در میان زنان جامعه امروز می‌باشد. اخیرا شیوع این بیماری افزایش یافته است (۱). از آنجا که تشخیص خوش‌خیم یا بدخیم بودن تومور در مراحل ابتدایی این بیماری امکان درمان و عمر طولانی مدت مبتلایان به آن را تضمین می‌نماید، متخصصین به دنبال روش‌های بهینه جهت بهبود تشخیص این تومور می‌باشند (۱-۳).

امروزه فرایندها و آزمایشات فراوانی جهت شناسایی سرطان پستان مانند گرفتن بیوپسی وجود دارد. همچنین جهت گرفتن بیوپسی به منظور تشخیص سرطان، روش-های متعددی نظیر استخراج سوزن ریز (FNA)، بیوپسی سوزن درونی، بیوپسی به کمک خلا و بیوپسی عمل باز و استخراج سوزن ریز وجود دارد (۴). علاوه بر اینها جهت تشخیص نیز استفاده از جداکننده‌ها (classifier) متداول می‌باشد. بی‌شک ارزیابی اطلاعات مستخرج از بیمار و تصمیمات متخصصین عوامل مهم در فرایند تشخیص می‌باشند، لیکن سیستم‌های تخصصی و استفاده از روش‌های مبتنی بر هوش مصنوعی به میزان چشمگیری به متخصصین کمک می‌نماید. این روش‌ها خطاهای متخصصین و زمان مورد نیاز آنان را به حداقل می‌رساند (۵).

یکی از عملیات اصلی جهت تشخیص، استفاده و استخراج اطلاعات مفید از داده‌های تشخیصی ماقبل می‌باشد. روش یادگیری ماشین رایانه‌ها را قادر به یادگیری از تجارب، الگوها و مثال‌های قبل می‌سازد (۲) از این رو، استفاده از این روش‌ها رو به رشد می‌باشد. در مورد سرطان پستان مشکل پیش رو تشخیص خوش‌خیمی یا بدخیمی تومور با استفاده از خواص نمونه می‌باشد. روش‌های زیادی تاکنون برای حل این مساله بکار گرفته شده است که اکثرا از داده‌های بانک اطلاعاتی ویسکانسین استفاده کرده‌اند (۱،۶،۷). کوینلان و همکاران از درخت تصمیم c4.5 استفاده نموده و به دقت تفکیک ۹۴،۷۴٪ رسید (۸). پس از آن تلاش محققین و استفاده از روش‌های مختلف از جمله تکنیک‌های فازی نظارت شده، فازی ژنتیک وآنیلینگ دقت را به ۹۸/۸٪ رساندند (۶-۷). در سال ۲۰۰۷ از تفکیک کننده‌های متفاوتی نظیر شبکه عصبی بردار پشتیبان، شبکه عصبی احتمالاتی و اختصاص منبع فازی استفاده گردید و دقت‌های ۹۹/۵۴، ۹۸/۶۱ و ۹۸/۵۱ حاصل شد (۸-۹). از سال ۲۰۰۷ تاکنون نیز روش‌های متعددی ارایه گردیده و تلاش برای بهبود تشخیص یا کاهش هزینه محاسبات همچنان ادامه دارد. در این میان، بهینه سازهای الگوریتم تکاملی نیز به دلیل کارایی مطلوبشان مورد توجه محققین قرار گرفته است (۹). از جمله ژو و همکاران که در سال ۲۰۱۴ با ارایه PSO بهینه شده توانست به دقت ۹۴/۷۴٪ دست یابد (۱۰).

در این مطالعه، عملکرد شبکه عصبی با استفاده از جغرافیای زیستی بهینه می‌گردد. بدین منظور با استفاده از آنالیز اجزای اصلی در مرحله آماده‌سازی و بروز رسانی همزمان وزن‌ها در طی فرایند، سعی بر بهینه سازی دسته‌بندی داده‌ها به عنوان خوش‌خیم یا بدخیم می‌شود. جهت ارزیابی الگوریتم ارایه شده از داده‌های بانک اطلاعاتی ویسکانسین استفاده می‌گردد (۱۲).

مواد و روش‌ها

در ابتدای این بخش بانک داده ویسکانسین بررسی و مختصری از شبکه عصبی پرسپترون بیان می‌شود. سپس روش پیشنهادی که به صورت آنالیز اجزای اصلی و الگوریتم بهینه سازی جغرافیای زیستی است، بیان می‌گردد. در انتهای بخش نیز نحوه آماده‌سازی شبکه عصبی بهینه شده بر پایه جغرافیای زیستی و طریقه آموزش آن بررسی می‌شود.

بانک داده سرطان پستان ویسکانسین: در این مطالعه، آزمایش روی بانک اطلاعاتی ویسکانسین (WBCD) که از مخزن یادگیری ماشین UCI اقتباس شده است، انجام گرفت (۱۱). بانک اطلاعاتی سرطان پستان ویسکانسین از ۶۹۹ نمونه که از استخراج سوزن ریز بافت پستان انسان گرفته شده، تشکیل شده است. اطلاعات استخراج سوزن ریز شامل ۹ ویژگی و کلاس (خوش‌خیم یا بدخیم) متناظر با هر نمونه می‌باشد. مقدار هر ویژگی که در جدول ۱ نشان داده شده است دارای مقدار صحیح بین ۱ تا ۱۰ می‌باشد. افزایش این رقم به معنای وخیم‌تر شدن وضعیت است، به طوری که مقدار ۱۰ به معنای وضعیت بسیار غیرعادی است. از ۶۹۹ نمونه، ۱۶ نمونه ناقص بوده و حذف گردید. همچنین، از نرمال‌سازی جهت پیش پردازش داده‌ها استفاده گردید. بنابراین آزمایش در نهایت بر روی ۶۸۳ نمونه باقی مانده که ۴۴۴ نمونه از آن متعلق به کلاس خوش‌خیم و ۲۳۹ نمونه

متعلق به کلاس بدخیم می‌باشد، انجام گردید. از آنجایی که غده سرطانی می‌تواند خوش‌خیم (مضر نمی‌باشد) و بدخیم (پتانسیل مضر بودن دارد) باشد، هدف الگوریتم ارایه شده جداسازی صحیح نمونه‌ها به عنوان خوش‌خیم یا بدخیم است.

جدول ۱: توضیح مشخصات بانک اطلاعاتی ویسکانسین

شناسه	ویژگی	دامنه
A1	ضخامت توده	۱-۱۰
A2	یکنواختی اندازه سلول	۱-۱۰
A3	یکنواختی شکل سلول	۱-۱۰
A4	چسبندگی حاشیه ای	۱-۱۰
A5	اندازه سلول مخاطی منفرد	۱-۱۰
A6	هسته‌های بی تحرک	۱-۱۰
A7	کروماتین بلاند	۱-۱۰
A8	هسته‌های طبیعی	۱-۱۰
A9	مایتوزها	۱-۱۰

شبکه عصبی (پرسپترون): شبکه عصبی مصنوعی[1] (ANN) (۱۲) تکنیک محبوبی از یادگیری ماشین است که از مغز انسان شامل شبکه عصبی بیولوژیکی الگو گرفته شده است و قابلیت یادگیری مناسبی دارد. شبکه‌های عصبی پیشخور (فیدفوروارد) نوع متداول شبکه عصبی بوده که در آن هر نورون مصنوعی وزنی دارد که به آن نسبت داده شده است. انتساب وزن‌ها با استفاده از ورودی‌هایی که از نورون‌های لایه قبل است و خروجی جهت پردازش به لایه بعدی منتقل می‌شود. نوع مهمی از شبکه‌های عصبی پیشخور، پرسپترون چند لایه[2] (MLP) است (۱۳).

الگوریتم پس انتشار[3] (BPA) یکی از تکنیک‌های متداول آموزش MLP بوده که با هر تکرار وزن‌های بین نورون‌ها را تغییر می‌دهد به طوری که خطا کمینه گردد. این مدل در یادگیری الگو نیز بسیار موفق بوده و می‌تواند روندهای جدید خو گیرد اما نقطه ضعفش سرعت کم همگرایی و واقع شدن در بهینه محلی است (۱۴).

مشکل دیگر BPA تصمیم‌گیری در چگونگی ساختار، تعداد لایه‌ها و تعداد نورون‌ها هر لایه است. این انتخاب نقش اساسی در عملکرد شبکه عصبی دارد و نتایج با تغییر اندک هر یک از پارامترهای بیان شده به طور چشمگیری تغییر کند. معماری متفاوت شبکه عصبی برای مسایل مختلف، نتایج مختلفی به همراه دارد. مشکلات کاربردی مستلزم ساختاریست که پاسخی بهینه داشته باشد. با این وجود، رسیدن به معماری بهینه شبکه عصبی به روش سعی و خطا کار سنگینی است.

پس از بهینه سازی شبکه عصبی توسط الگوریتم ژنتیک (۱۵) تلاش‌های فراوانی جهت بهبود عملکرد این شبکه‌ها انجام شد (۱۶، ۱۷). به هرحال هیچ کدام از اصلاحات قادر به عملکرد مطلوب برای همه مسایل نیست. بنابراین، جستجو برای سرعت بخشی به همگرایی و یافتن ساختار بهینه همچنان ادامه دارد. به همین دلیل، ما مدل شبکه عصبی بهینه شده بر پایه جغرافیای زیستی که ساختار بهینه و وزن‌های معماری شبکه عصبی را می‌یابد، معرفی می‌نماییم. در این روش وابستگی داده‌های ورودی شبکه عصبی به کمک الگوریتم آنالیز اجزای اصلی از بین رفته و داده‌های مستقل به عنوان ورودی شبکه عصبی در نظر گرفته می‌شود. در بخش‌های بعدی آنالیز اجزای اصلی و الگوریتم بهینه سازی جغرافیای زیستی مورد بحث قرار می‌گیرد.

آنالیز اجزای اصلی[4] (PCA): یکی از مشکلات اساسی در امر شبیه سازی، یافتن روش‌هایی مانند آنالیز مود نرمال (۱۸، ۱۹) یا آنالیز اجزای اصلی(۲۰) برای کاهش ابعاد داده‌های ورودی می‌باشد. آنالیز اجزای اصلی یکی از روش‌هایی است که همواره برای این منظور بکار برده می‌شود (۲۰، ۲۱). علاوه بر این، این روش در انتخاب ویژگی برای داده‌های سرطان پستان موفق عمل کرده است (۲۲).

این تکنیک سعی به پیدا کردن تبدیل خطی $v=Wu$ (در اینجا u بردار مشاهدات می‌باشد) می‌کند بطوری که واریانس بدست آمده بیشینه باشد که منجر به حل یک مساله مقدار ویژه متقارن می‌شود. بردارهای سطری W مطابق با بردارهای ویژه نرمال ماتریس کواریانس داده‌ها است. یکی از روش‌های ساده برای حل PCA استفاده از تجزیه مقدار منفرد (svd) است. اگر ماتریس کوواریانس $Ru = E[uu^T]$. باشد. SVD ماتریس Ru به شکل

[1] Artificial Neural Network
[2] Multiple Layer Perceptron
[3] Back Propagation Algorithm
[4] Principal Component Analysis

$Ru = U_u\ D_u\ U_u^T$ است بطوری که ماتریس بردارهای ویژه و D ماتریس قطری که عناصر قطریش مطابق با مقادیر ویژه Ru می‌باشد. تبدیل خطی w برای PCA با رابطه $W= U_u^T$ نشان داده می‌شود. جهت کاهش ابعاد بردار ستونی غالب P در U_u انتخاب می‌گردد که همان بردارهای ویژه مرتبط با بزرگترین مقدار ویژه p جهت ساخت تبدیل خطی W می‌باشد (۲۳). از داده‌های سرطان پستان ویسکانسین WBCD که توسط ولبرگ ارایه گردید جهت امتحان این روش پیشنهادی استفاده شد. پس از کاهش ابعاد وعدم وابستگی، این بردارهای ویژگی جهت تفکیک وارد شبکه عصبی پرسپترون چند لایه می‌شوند.

الگوریتم بهینه سازی جغرافیای زیستی: الگوریتم بهینه سازی جغرافیای زیستی[5] BBO یک الگوریتم تکاملی بر پایه جمعیت جانوران موجود در یک زیستگاه جغرافیای است (۲۴). این الگوریتم بر پایه پدیده مهاجرت جانوران به زیستگاه‌های مختلف بنا شده است. به طور کلی زیستگاه‌هایی که مکان مناسبی برای گونه‌های جغرافیایی جهت اسکان هستند، دارای شاخص تناسب زیستگاه[6] (HSI) بالا هستند. این شاخص توسط متغیرهای سکنی که متغیر شاخص تناسب[7] (SIV) نام دارد، تعیین می‌گردد.

زیستگاه با شاخص تناسب زیستگاهی بالا، دارای گونه‌هایی می‌باشند که به زیستگاه‌های اطراف مهاجرت می‌کنند. زیستگاه با HSI بالا دارای نرخ مهاجرت‌پذیری کمی هستند، چرا که از قبل توسط گونه‌های دیگر اشغال شده‌اند و نمی‌توانند پذیرای گونه‌های جدید باشند. از طرفی دیگر، زیستگاه با HSI پایین به دلیل جمعیت اندک خود دارای نرخ مهاجرت‌پذیری بالایی هستند. مهاجرت‌پذیری گونه‌های جدید به زیستگاه‌های دارای HSIپایین‌تر می‌تواند باعث افزایش HSI آن منطقه شود، زیرا مناسب بودن یک مکان، متناسب با تنوع جغرافیایی آن است.

اساس کار الگوریتم‌های تکاملی مانند ژنتیک، ایجاد جمعیت اولیه و سپس استفاده از عملگرهای خاص مانند ترکیب و جهشی در آنها به منظور تولید جمعیت بهتر یا دست‌یابی به پاسخ دقیق‌تر می‌باشد. تعداد جمعیت جدید برابر با اختلاف جمعیت اولیه با جمعیتی است که توسط نرخ نگه داری نگاه داشته می‌شود (جدول۲). از این رو، در الگوریتم جغرافیای زیستی نیز به منظور ایجاد تغییرات مطلوب در روند تولید جمعیت نسل‌ها یا همان پاسخ‌ها، از دو عملگر مهاجرت و جهش استفاده می‌شود. این پاسخ‌ها توسط تابع برازش[8] که در الگوریتم بهینه سازی جغرافیای زیستی همان HIS می‌باشد، مورد ارزیابی قرار می‌گیرند (۲۴).

عملگر مهاجرت: پس از ایجاد پاسخ‌های اولیه از روش‌هایی برای تعیین میزان مطلوب بودن و طبقه‌بندی آنها استفاده می‌شود. پاسخ‌های مطلوب دارای (HSI) بالا به معنای زیستگاه با گونه‌های زیاد و پاسخ‌های ضعیف دارای HSI پایین به معنای زیستگاه با گونه‌های کم هستند. هر زیستگاه در الگوریتم بهینه سازی جغرافیای زیستی دارای نرخ مهاجرت (λ) و نرخ مهاجرت‌پذیری (μ)برای به اشتراک گذاری اطلاعات به صورت احتمالی بین راه حل‌ها استفاده شده، با روابط زیر محاسبه می‌شوند:

$$\lambda_i = I(1 - \frac{k(i)}{n}) \qquad (۱)$$

$$\mu_i = E(\frac{k(i)}{n}) \qquad (۲)$$

که در آن I و E به ترتیب بیشترین مقدار نرخ مهاجرت و مهاجرت‌پذیری است که پاسخ‌ها می‌توانند داشته باشند و k(i) نشان‌دهنده تعداد گونه‌ها در زیستگاه i ام است. این مقدار بین 1 تا n می‌باشد. n نیز تعداد اعضای جمعیت است. هر پاسخ، با احتمالی خاص برای اصلاح پاسخ دیگر به کار می‌رود. زمانی که یک پاسخ برای اصلاح انتخاب می‌گردد، از نرخ مهاجرت‌پذیری (λ) آن استفاده می‌شود تا تعیین شود هر یک از SIV های موجود برای پاسخ باید اصلاح شود یا خیر.

زمانی یک SIV موجود درپاسخ Si برای اصلاح انتخاب شد، با کمک نرخ مهاجرت (μ) به شکل احتمالی، تصمیم می‌گیریم که کدام یک از پاسخ‌ها، باید باعث مهاجرت یک SIV، که به صورت تصادفی انتخاب شده است به پاسخ Si گردد.

[5] Biogeography Based Optimization
[6] High Suitable Index
[7] Suitability Index Variable
[8] Fitness

جهش: تحولات ناگهانی می‌تواند مقدار HSI یک زیستگاه را تغییر دهد. همچنین می‌تواند باعث شوند که تعداد گونه‌ها با مقدار متعادل خود متفاوت باشند. این امر را به عنوان جهش SIV در BBO مدل می‌کنیم از احتمال تعداد گونه‌های موجود در زیستگاه برای مشخص کردن نرخ جهش استفاده می‌شود.

$$m_s = m_{max}(1 - \frac{p_s}{p_{max}}) \quad (۳)$$

که در آن m max بیشترین مقدار نرخ جهش بوده که توسط کاربرمعرفی می‌شود. Ps احتمال اینکه زیستگاه دقیقا دارای s گونه باشد (۲۴). این الگوی جهش منجر به افزایش تنوع در جمعیت می‌شود. پارامترهای BBO برای بهینه سازی شبکه عصبی این مطالعه در جدول ۲ آورده شده است.

جدول ۲: پارامترهای انتخاب شده بهینه ساز جغرافیای زیستی

پارامترها	مقادیر
اندازه جمعیت	۲۰
بیشینه مقدار تکرار	۳۰
نرخ نگهداری	۰/۴
نرخ جهش	۰/۱

شبکه عصبی بهینه شده بر پایه جغرافیای زیستی (BBONN)[9]: الگوریتم BBO در خروجی شبکه‌های عصبی پرسپترون مطابق الگوی ارایه شده در شکل ۱ استفاده گردید. جهت ارزیابی تابع فیتنس HSI خطای مجذور میانگین MSE[10] برای تمامی نمونه‌های یک زیستگاه محاسبه گردید. MSE به صورت معادله ۴ نمایش داده می‌شود:

$$MSE = \frac{1}{k}\sum_{i=1}^{k}(y - \hat{y})^2 \quad (۴)$$

بطوری که در آن y خروجی واقعی و$\hat{y}$ خروجی تخمین زده شده وk تعداد نهایی نمونه‌ها در مجموعه آموزش می‌باشد. شکل ۱ نشان می‌دهد که BBO در ابتدا کاندیدهای تصادفی که به طبع آن وزن‌ها و بایاس رندوم‌زا شامل می‌شود. این الگوریتم سپس MSE را برای تمامی شبکه‌های پرسپترون ایجاد شده در هنگام دسته‌بندی محاسبه می‌کند. خطای MSE نشان می‌دهد که کدام شبکه نرونی پرسپترون مناسب‌تر است.

بر اساس قوانین بیان شده، الگوریتم BBO مجموعه‌ای از شبکه‌های پرسپترون جدیدی را با توجه به بهترین شبکه‌ای که تاکنون به وجود آمده است، به وجود می‌آورد. فرایند محاسبه MSE و بهبود شبکه پرسپترون آنقدر ادامه می‌یابد تا شرط پایان ارضا شود. این شرط می‌تواند یک حد آستانه یا بیشینه تکرارها باشد. باید توجه داشت که MSE میانگین در هنگام دسته‌بندی نمونه‌های آموزش در بانک داده برای هر شبکه پرسپترون برای آموزش بر اساس BBO محاسبه می‌گردد.

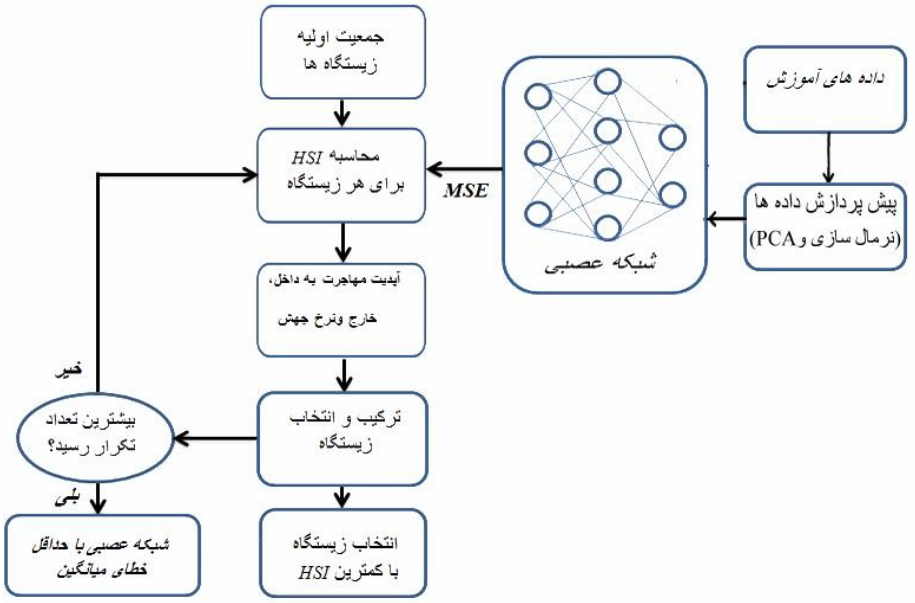
یافته‌ها

روش BBONN ارایه شده به عنوان تفکیک کننده بر روی بانک داده‌های سرطان پستان ویسکانسین (WBCD) اعمال گردید. عملیات با استفاده از پارامترهای جدول ۲ انجام شد. در زمینه یادگیری ماشین، روش معمول به این صورت است که بانک داده به دو زیر مجموعه مجزای داده‌های آموزش و داده‌های تست تقسیم‌بندی می‌شود. جهت محک عمومیت روش ارایه شده در این مطالعه و مقایسه کار با مطالعات موجود این زمینه، داده‌های آموزش و داده‌های تست را به چهار نسبت مختلف تقسیم‌بندی کردیم. نسبت۵۰-۵۰، که در آن نیمی از داده‌ها را برای آموزش شبکه و نیمی دیگر را جهت تست انتخاب می‌شوند. همچنین از نسبت‌های ۶۰-۴۰، ۵۰-۵۰ و ۷۰-۳۰ با اعتبار سنجی ضربدری ۱۰ گانه[11] جهت جلوگیری از اورفیتینگ[12] و نشان دادن اهمیت روش ارایه شده استفاده کردیم.

برای BBONN ما معماری شبکه‌ای را که در آن دو لایه پنهان که لایه اول شامل یک نورون و لایه دوم شامل دو نورون می‌باشد و همچنین یک لایه خروجی که دارای یک نورون بود، انتخاب نمودیم. بنابراین ساختار شبکه در بر دارنده یک لایه ورودی، یک لایه خروجی و دولایه پنهان (۱-۲-۱) می‌باشد.

تابع فعال سازی برای الگوریتم‌های مورد استفاده در این مطالعه سیگموید بوده و برای تمامی نورون‌ها اعمال می‌شود.

[9] Biogeography Based Optimized Neural Network
[10] Minimum square error
[11] 10 fold cross validation
[12] Over-fitting



شکل۱: فلوچارت الگوریتم پیشنهادی

دو معیار مختلف جهت متوقف سازی آموزش استفاده گردید: در حالت اول زمانی متوقف شد که متوسط مجذور خطا به ۰/۰۱ رسید و در حالت دیگر آموزش زمانی متوقف گردید که تعداد کل ارزیابی فیتنس به ۴۰۰۰ و تعداد کل تکرارها (epocs) به ۲۰۰۰ برسد. تعداد تکرارها، تعداد دفعاتی است که نمونه‌های آموزش به شبکه معرفی می‌گردند. انحراف معیار ماکزیمم، میانگین و استاندارد برای نتایج با ۵۰ تکرار برای نسبت‌های تقسیم‌بندی ۵۰-۵۰، ۶۰-۴۰و۷۰-۳۰ ارایه گردیده است.

با توجه به نتایج به دست آمده و به منظور مقایسه و ارزیابی عملکرد مدل BBONN پیشنهادی، ما دقت تفکیک، حساسیت و ویژگی را با توجه به مقادیر محاسبه شده تعریف و محاسبه می‌نماییم. فرمول‌بندی به‌صورت زیر (جدول ۳) می‌باشد.

جدول۳: روش محاسبه دقت تفکیک،حساسیت و ویژگی

عنوان	رابطه	عنوان	رابطه
حساسیت	$Sen = \frac{TP}{TP+FN} * 100$	کسر مثبت نادرست	$FPF = \frac{FP}{TN+FP} * 100$
ویژگی	$Spe = \frac{TN}{TN+FP} * 100$	کسر منفی نادرست	$FNF = \frac{FN}{TP+FN} * 100$
دقت	$Acc = \frac{TP+TN}{TP+TN+FP+FN} * 100$	کسر موارد نابجا	$Misclass = \frac{FP+FN}{TP+TN+FP+FN} * 100$

FP: مثبت نادرست، تعداد نمونه‌های خوش‌خیمی که به اشتباه بدخیم شناخته شده‌اند. TN: منفی درست، بیانگر تعداد نمونه‌های خوش‌خیمی که به درستی شناخته شده‌اند. TP: مثبت درست، تعداد نمونه‌های بدخیمی که به درستی شناخته شده‌اند. FN: منفی نادرست، تعداد نمونه‌های بدخیمی که به اشتباه خوش‌خیم شناخته شده-اند.

مقایسه با حالات غیر بهینه: روش ارایه شده در این مطالعه با حالات غیر بهینه، که شامل حالت عادی و حالتی که از روش آنالیز اجزای اصلی نیز کمک گرفته شود، بر روی داده‌های سرطان پستان ویسکانسین انجام شده و نهایتا نتایج مقایسه گردید. دقت بدست آمده در حالت عادی یعنی بدون بهینه سازی، با بکارگیری آنالیز اجزای مستقل و بکارگیری اجزای مستقل و بهینه ساز زیستی در جدول ۴ آورده شده است. نتایج این جدول نشان می‌دهد در حالی که نسبت آموزش ۵۰-۵۰ می‌باشد استفاده از آنالیز اجزای مستقل تاثیر بسزایی در افزایش دقت دارد.

از آنجایی که در برخی موارد دقت به تنهایی بیان‌گر کارایی روش استفاده شده نمی‌باشد مقادیر ویژگی و حساسیت نیز برای حالات عادی، با بکارگیری آنالیز اجزای مستقل و بکارگیری توامان اجزای مستقل و بهینه‌ساز زیستی محاسبه گردید (جدول ۵ و ۶).

منحنی ROC برای حالت عادی، با بکارگیری آنالیز اجزای مستقل و بکارگیری اجزای مستقل و بهینه ساز زیستی به طور همزمان برای نسبت داده ۷۰-۳۰ به ترتیب در شکل‌های ۲ الف، ب و ج آورده شده است. سطح زیر منحنی ROC یا (AUC) توسط نرم افزار متلب و به روش ذوزنقه محاسبه گردیده است. برای حالت عادی ۰/۷۸، آنالیز اجزای مستقل ۰/۹۷ و بکارگیری اجزای مستقل و بهینه ساز زیستی به طور همزمان ۱ می‌باشد. به طور مشابه، مقدار AUC برای داده‌های ۵۰-۵۰ و ۶۰-۴۰ نیز برای اجزای مستقل و بهینه‌ساز زیستی به طور همزمان به ترتیب ۰/۹۹۱ و ۰/۹۹۷ می‌باشد.

جدول ۴: دقت بدست آمده نسبت اموزش به تست ۵۰ به ۵۰ ، ۶۰-۴۰ و ۷۰-۳۰ به ترتیب برای حالت عادی با بکارگیری آنالیز اجزای مستقل و بکارگیری اجزای مستقل و بهینه ساز زیستی به طور همزمان

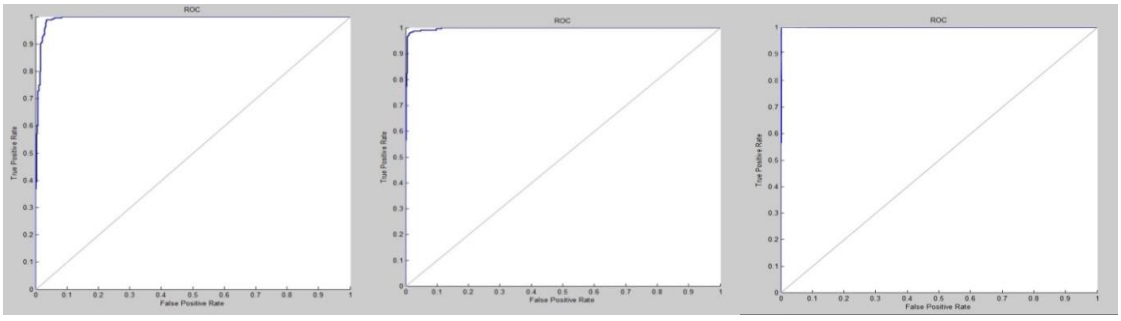
نسبت آموزش به تست	حالت عادی	آنالیز اجزای مستقل (PCA)	اجزای مستقل و جغرافیای زیستی (PCA+BBO)
۵۰-۵۰	٪۶۵	٪۹۷/۷	٪۹۸/۴
۴۰-۶۰	٪۹۷/۸	٪۹۸/۲	٪۹۸/۷
۳۰-۷۰	٪۹۸/۵	٪۹۹	٪۱۰۰

جدول ۵: ویژگی بدست آمده نسبت آموزش به تست ۵۰ به ۵۰ ، ۶۰-۴۰ و ۷۰-۳۰ به ترتیب برای حالت عادی با بکارگیری آنالیز اجزای مستقل و بکارگیری اجزای مستقل و بهینه ساز زیستی به طور همزمان

نسبت آموزش به تست	حالت عادی	آنالیز اجزای مستقل (PCA)	اجزای مستقل و جغرافیای زیستی (PCA+BBO)
۵۰-۵۰	۱	۰/۹۷	۰/۹۷
۴۰-۶۰	۰/۹۹	۰/۹۸	۰/۹۸
۳۰-۷۰	۰/۹۹	۰/۹۸	۱

جدول ۶: حساسیت بدست آمده نسبت آموزش به تست ۵۰ به ۵۰ ، ۶۰-۴۰ و ۷۰-۳۰ به ترتیب برای حالت عادی با بکارگیری آنالیز اجزای مستقل و بکارگیری اجزای مستقل و بهینه ساز زیستی به طور همزمان

نسبت آموزش به تست	حالت عادی	آنالیز اجزای مستقل (PCA)	اجزای مستقل و جغرافیای زیستی (PCA+BBO)
۵۰-۵۰	۰/۹۵	۰/۹۶	۰/۹۶
۴۰-۶۰	۰/۹۷	۰/۹۶	۰/۹۶
۳۰-۷۰	۰/۹۸	۰/۹۷	۱



شکل۲: سطح زیر نمودار ROC برای نسبت‌های آموزش به تست ۵۰-۵۰، ۶۰-۴۰ و ۷۰-۳۰

بحث

مطالعه فعلی با هدف بهینه سازی و ارزیابی یک سیستم مبتنی بر شبکه عصبی جهت کمک به تشخیص پزشک در تعیین نوع توده‌های سرطان پستان انجام شد. سیستم ارایه شده شده در این مطالعه در تشخیص توده‌های خوشخیم و بدخیم موفق بود و دسته‌بندی را با دقت ۱۰۰٪ انجام داد. حسن روش پیشنهادی BONN استفاده از کاهش ویژگی به روش آنالیز جزء اصلی و انتخاب مناسب بهینه ساز تکاملی بود که نتایج حاصل از آن حاکی از سرعت بیشتر و تعمیم‌پذیری بهتر در عین افزایش دقت نسبت به دیگر موارد پیاده سازی شده در این موضوع بود. بر این اساس، این روش می‌تواند ابزار بسیار مناسبی جهت کمک به پزشکان برای تشخیص بیماری باشد و یا به عنوان نظر دوم برای تشخیص نهایی مورد استفاده قرار گیرد. استفاده از چنین روش‌های دقیق و سریع، امید عملی کردن سامانه تشخیص هوشمند سرطان پستان را بیشتر می‌کند.

نتایج شبیه سازی نشان داد سیستم پیشنهادی در این مقاله بر روی مجموعه داده بیماران مبتلا به سرطان پستان بیمارستان ویسکانسین به دقت ۱۰۰٪ رسیده است که بالاتر از تحقیقات مشابه بر روی این مجموعه داده بوده است. علاوه بر این باید به مقدار پارامتر منفی کاذب نیز توجه داشت. درصد این پارامتر در مدل‌های پیش بینی در حوزه پزشکی بسیار اهمیت دارد چون فرد بیمار به اشتباه، سالم در نظر گرفته می‌شود که می‌تواند عواقب بسیار خطرناکی داشته باشد. در مدل پیش بینی ارایه شده در این تحقیق این مقدار برای دسته‌بندی ۷۰-۳۰ صفر بود، که از دیگر مزیت‌های روش پیشنهادی است.

در زمینه دقت مدل‌های ایجاد شده، مقایسه نتایج مطالعات اخیر بر روی پایگاه داده ویسکانسین با نتایج این مطالعه حایز اهمیت است. مالمیر و همکاران (۲۵) با استفاده روش بهینه سازی رقابت استعماری دقت شبکه عصبی را به دقت ۹۷/۷٪ رسیدند. کویونکو و سایان (۲۶) نیز با استفاده از شبکه عصبی شعاعی بر روی مجموعه داده بیمارستان ویسکانسین به دقت ۹۶٪ رسیدند. اوبیلی و همکاران (۲۷) با بکارگیری ماشین بردار پشتیبان برای تشخیص سرطان به ۹۹/۵٪ رسیدند. در مقایسه با نتایج به دست آمده در مطالعات ذکر شده پیشین، در این پژوهش با استفاده از کاهش ویژگی به روش آنالیز جزء اصلی به شناسایی کامل دست یافت (جدول ۴). استفاده از تعداد کمتر ویژگی‌ها باعث بهبود سرعت کارکرد سیستم‌های برخط می‌شود.

از سوی دیگر سلطانی سروستانی و همکاران (۲۸) روش‌های مختلفی جهت تشخیص خوش‌خیم یا بدخیم بودن سرطان پستان با استفاده از شبکه‌های عصبی مختلف ارایه کردند و متوسط مربع خطا در هر شبکه را باهم مقایسه کردند که از میان شبکه‌های عصبی رقابتی و پایه شعاعی، شبکه عصبی پایه شعاعی به کمترین متوسط خطا و در نتیجه دقت بهتری دست یافت. در مطالعه دیگری مدحت محمد احمد و همکاران (۲۹) از ماشین بردار پشتیبان استفاده نمودند و دریافتند که این روش دارای سطح زیر منحنی بیشتری است. نتایج مطالعات ذکرشده محققان بر روی این مجموعه داده در مقایسه با روش پیشنهادی این مقاله، نشان می‌دهد که این روش از این حیث نیز برکارهای گذشته در این حوزه برتری داشته و سطح زیر منحنی را به حداکثر رسانیده است (شکل ۲). استفاده از چنین روش‌های دقیق و سریع، امید عملی کردن سامانه تشخیص هوشمند سرطان پستان را بیشتر می‌کند.

از محدودیت‌های این پژوهش می‌توان به ۱۶ رکورد حذف شده از مجموعه داده اشاره کرد که می‌تواند بر نتایج به دست آمده از این پژوهش مؤثر باشد. به روز نبودن مجموعه داده‌ها از دیگر محدودیت‌هایی است که با توجه به آمار سالانه مبتلایان به این بیماری، بهتر است مدل سازی با داده‌های واقعی و به روز که تعداد کافی از بیماران را در بر گیرد، انجام پذیرد و نتایج ارزیابی شود. به علاوه، نتایج این پژوهش فقط برای پایگاه داده مورد مطالعه (مگر به شرط توسعه) معتبر می‌باشد.

شایان ذکر است که این مدل، برای تشخیص سرطان پستان که یک مساله دو کلاسه است، پیشنهاد شد. بنابراین رفتار آن در سایر مسایل مرتبط با طبقه بندی چند کلاسه نیاز به بررسی دارد. علاوه بر این، مدل پیشنهادی تنها بر روی داده‌های مقداری اعمال شد که می توان رفتار این الگوریتم را بر روی تصاویر و سیگنال‌های مورد مطالعه این حوزه نیز قرار داد. همچنین از آنجایی که موجود نبودن داده‌های بالینی داخلی جامع یکی از محدودیت‌های مطالعه حاضر می‌باشد، برای مطالعات آینده پیشنهاد می‌گردد با بومی سازی این

سیستم، شبکه را با مجموعه داده‌های بیمارستان‌های کشورمان آموزش داد و به پیش‌بینی این بیماری پرداخت.

نتیجه‌گیری

در این مقاله از الگوریتم جغرافیای زیستی جهت بهینه سازی خروجی شبکه عصبی برای طبقه‌بندی نوع سرطان پستان به دو دسته خوش‌خیم و بدخیم استفاده شد. ابتدا داده‌های مورد استفاده با نرمال سازی و استفاده از آنالیز اجزای مستقل پیش پردازش شدند. سپس شبکه با نمونه‌های آموزشی و اعتبار سنج، آموزش داده و اعتبار سنجی شد و در انتها با نمونه‌های آزمون آزموده شد. از دلایل بالا بودن حساسیت و ویژگی در مقاله حاضر می‌توان به پیش پردازش داده‌های ورودی همان‌طور که پیش‌تر شرح داده شد و انتخاب مناسب بهینه ساز شبکه عصبی برای این منظور اشاره کرد. در واقع روش پیشنهادی در این مقاله به خاطر سرعت و دقت زیاد و تعمیم پذیری خوب آن نسبت به دیگر روش‌ها برتر است. مطالعاتی از این دست برای مطالعات بعدی نیز می‌توانند بررسی و استفاده شوند و از طرفی به دلیل کم هزینه بودن و سرعت بالای انجام فرآیند بسیار به صرفه خواهند بود. پیشنهاد می‌شود بر این اساس نرم افزاری طراحی شده و جهت کمک به پزشکان مورد استفاده قرار گیرد.

References

1. Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst Appl 2009; 36(2):3240–7.
2. R Sheikhpour MAS. Selection of Relevant and Effective Features in Detection of Breast Cancer using Parametric Learning Methods [Internet]. IJBD YR. 2015. p. 16–23. Available from: http://ijbd.ir/article-1-435-fa.html
3. Maglogiannis I, Zafiropoulos E, Anagnostopoulos I. An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers. Appl Intell 2009;30(1):24–36.
4. Alipoor M, Haddadnia J. An accurate intelligent breast cancer diagnosis system 2009; 2(5):33-40
5. Xiong X, Kim Y, Baek Y, Rhee DW, Kim S. Analysis of Breast Cancer Using Data Mining & Statistical Techniques 2005;3–8.
6. Aličković E, Subasi A. Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Comput Appl 2017;28(4):753–63.
7. Durai SG, Ganesh SH. A Novel Iterative Linear Regression Perceptron Classifier for Breast Cancer Prediction. Int J Comput Appl 2017;167(12).
8. Quinlan JR. Improved use of continuous attributes in C4. 5. J Artif Intell Res 1996;4:77–90.
9. Koyuncu H, Ceylan R. Artificial neural network based on rotation forest for biomedical pattern classification. In: Telecommunications and Signal Processing (TSP), 2013 36th International Conference on IEEE 2013; 581–5.
10. Xue B, Zhang M, Browne WN. Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms. Appl Soft Comput 2014;18:261–76.
11. Blake CL, Merz CJ. UCI repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Sciences 1998. 2005;572-7.
12. Denton JW, Hung MS, Osyk BA. A neural network approach to the classification problem. Expert Syst Appl 1990;1(4):417–24.
13. Faris H, Aljarah I, Mirjalili S. Training feedforward neural networks using multi-verse optimizer for binary classification problems. Appl Intell [Internet]. 2016;(March). Available from: http://dx.doi.org/10.1007/s10489-016-0767-1
14. Gupta JND, Sexton RS. Comparing backpropagation with a genetic algorithm for neural network training. Omega 1999;27(6):679–84.

15. Koza JR, Rice JP. Automatic programming of robots using genetic programming. In AAAI 1992; 194–207.

16. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE, et al. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer 1997;79(4):857–62.

17. Abbass HA. An evolutionary artificial neural networks approach for breast cancer diagnosis. Artif Intell Med 2002;25(3):265–81.

18. Teodoro ML, Phillips Jr GN, Kavraki LE. Understanding protein flexibility through dimensionality reduction. Journal of Computational Biology. 2003 Jun 1;10(3-4):617-34

19. Toussi CA, Soheilifard R. A better prediction of conformational changes of proteins using minimally connected network models. Phys Biol 2017;13(6):66013.

20. Dou Y, Geng X, Gao H, Yang J, Zheng X, Wang J. Sequence conservation in the prediction of catalytic sites. Protein J 2011;30(4):229–39.

21. Hasan H, Tahir NM. Feature selection of breast cancer based on principal component analysis. In: Signal Processing and Its Applications (CSPA), 2010 6th International Colloquium on IEEE 2010; 1–4.

22. Lee H, Choi S. Pca+ hmm+ svm for eeg pattern classification. In: Signal Processing and Its Applications, 2003 Proceedings Seventh International Symposium on IEEE 2003; 541–4.

23. Simon D. Biogeography-based optimization. IEEE Trans Evol Comput 2008;12(6):702–13.