تشخیص هوشمند بیماری هپاتیت با استفاده از آنالیز اجزای اصلی و هم‌جوشی طبقه‌بندی‌کننده‌ها

سید جلال‌الدین موسوی راد*(Ph.D)، حسین ابراهیم‌پور کومله (Ph.D)
دانشگاه کاشان، دانشکده مهندسی برق و کامپیوتر، گروه مهندسی کامپیوتر

چکیده

سابقه و هدف: در سال‌های اخیر، بیماری هپاتیت در جهان بسیار شیوع پیدا کرده است. تشخیص صحیح بیماری هپاتیت کار ساده‌ای نمی‌باشد. هدف از این مقاله، ارائه یک سیستم هوشمند مبتنی بر تکنیک‌های یادگیری ماشین جهت تشخیص بیماری هپاتیت می‌باشد.

مواد و روش‌ها: الگوریتم پیشنهادی شامل سه مرحله اساسی می‌باشد: کاهش ابعاد، طبقه‌بندی و هم‌جوشی طبقه‌بندی‌کننده‌ها با یک‌دیگر. مجموعه داده‌ها از انباره داده‌های پایگاه داده‌ی UCI گرفته شده است. در ابتدا تمام داده‌ها نرمال شده‌اند. سپس با استفاده از آنالیز اجزای اساسی تعداد ویژگی‌ها به ۱۰ کاهش پیدا کرده است. در مرحله بعد از سه طبقه‌بندی‌کننده جهت مدل‌سازی داده‌ها استفاده گشته است. جهت بهبود کارایی و اطمینان بیش‌تر به نتایج سیستم، نتایج این سه طبقه‌بندی‌کننده با استفاده از رای‌گیری وزن‌دار با هم ترکیب شده است.

یافته‌ها: الگوریتم پیشنهادی توانست با استفاده از اعتبارسنجی منقطع ۱۰ لایه، دقت ۹۶/۳۲ را ارائه دهد که نسبت به کارهای مشابه نتیجه خوبی می‌باشد.

نتیجه‌گیری: با توجه به نتایج، سیستم پیشنهادی می‌تواند به عنوان یک همیار هوشمند جهت تشخیص نهایی پزشکان مورد استفاده قرار گیرد.

واژه‌های کلیدی: تشخیص هوشمند بیماری هپاتیت، یادگیری ماشین، هم‌جوشی طبقه‌بندی‌کننده‌ها، آنالیز اجزای اصلی.

مقدمه

در سال‌های اخیر، بیماری هپاتیت به طور گسترده‌ای در سطح جهان شیوع پیدا کرده است. هپاتیت به دو نوع حاد و مزمن تقسیم می‌شود و در اثر انواع ویروس‌ها، داروها، الکل و ... در کبد ایجاد می‌شود [۱]. انواع متعددی از ویروس‌ها می‌توانند در افراد ایجاد هپاتیت نمایند که سه نمونه از مهم‌ترین آن‌ها، هپاتیت نوع A، هپاتیت نوع B و هپاتیت نوع C می‌باشد. هپاتیت نوع A بیش‌تر در بچه‌ها دیده می‌شود و از طریق مواد غذایی یا آب آلوده به افراد سالم منتقل می‌گردد. کودک در ابتدا دچار حالتی شبیه سرماخوردگی و تب شده و با قطع تب دچار زردی در سفیدی چشم و پوست می‌گردد. یکی از خطرناک‌ترین ویروس‌های هپاتیت، هپاتیت نوع B می‌باشد. فردی که به هپاتیت B مبتلا می‌شود ممکن است احساس خستگی، ضعف، تهوع و بی‌اشتهایی نماید. عفونت حاد و مزمن هپاتیت B منجر به مرگ و میر ۵۰۰۰۰۰ تا ۱۲۰۰۰۰۰ نفر در سال می‌شود [۲]. هپاتیت C در اثر یک نوع دیگر از ویروس‌های هپاتیت به وجود می‌آید. هپاتیت C به طور عمده از طریق خون منتقل می‌شود.

* نویسنده مسئول، تلفن: ۰۹۳۵۴۶۲۶۳۳۴ نمابر: ۵۵۹۱۲۴۲۴ - ۰۳۱ Email: jalalmoosavirad@gmail.com

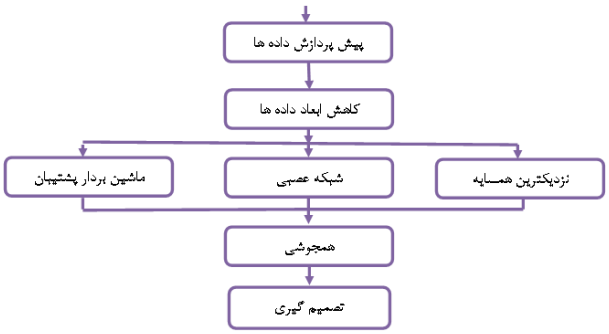
تاریخ دریافت: تاریخ پذیرش:

تشخیص صحیح بیماری هپاتیت برای یک پزشک کار مشکلی می‌باشد [۳]. یک پزشک با استفاده از تفسیر نتایج تست‌های انجام شده یک بیمار یا با مقایسه یک بیمار با علائم مشابه بیماران دیگر، به تشخیص این بیماری می‌پردازد [۴]. بر این اساس روش‌های هوشمند تشخیص این بیماری به یکی از موضوعات داغ در این حوزه تبدیل شده است. استفاده از الگوریتم‌های شناسایی الگو و یادگیری ماشین می‌تواند به عنوان یک ابزار برای تشخیص هوشمند این بیماری معرفی گردد. پلات [۵]، روشی مبتنی بر آنالیز اجزای اصلی و سیستم شناسایی ایمنی مصنوعی (Artificial immune recognition system, AIRS) برای تشخیص بیماری هپاتیت معرفی نموده است. در روش پیشنهادی، در ابتدا تعداد ویژگی‌ها به ۵ کاهش پیدا کرده و سپس ویژگی‌ها در فاصله بین ۰ تا ۱، نرمال شده‌اند. بعد از این مرحله، ویژگی‌ها به یک طبقه‌بندی‌کننده‌ی مبتنی بر AIRS داده شده است. دقت طبقه‌بندی این سیستم بر روی مجموعه داده‌های UCI [۶]، ۹۴/۱۲٪ گزارش شده است. دوگانتین و همکاران [۷]، برای تشخیص این بیماری، از ترکیب آنالیز جداکننده خطی و شبکه وفقی بر پایه‌ی سیستم استنتاج فازی (LDA-ANFIS) استفاده نموده‌اند و به دقت ۹۴/۱۶ بر روی مجموعه داده‌های UCI رسیدند. در کار دیگری، یک روش ترکیبی با نام PCA-LSSVM معرفی شده است [۴] که در ابتدا با استفاده از آنالیز اجزای اصلی تعداد ویژگی‌ها به ۱۰ کاهش پیدا کرده و ویژگی‌های جدیدی ساخته شده است، سپس از ماشین بردار پشتیبان جهت طبقه‌بندی استفاده کرده است. دقت الگوریتم پیشنهادی، با استفاده از پارامترهای مختلف به دست آمده و به دقت ماکزیمم ۹۶/۱۲ رسیده است. ماشین‌بردار پشتیبان در کار دیگری نیز استفاده شده است [۸]. نویسندگان در این کار با استفاده از آنالیز جداکننده فیشر و ماشین‌بردار پشتیبان دقت ۹۶/۷۷٪ را گزارش نموده‌اند.

طبقه‌بندی‌کننده‌های گوناگون، دارای محدودیت‌های مختلفی می‌باشند که باعث شده تا توسعه یک طبقه‌بندی‌کننده جدید با کارایی بالاتر کار مشکلی به نظر برسد. یک راه‌کار جهت افزایش کارایی طبقه‌بندی‌کننده‌ها، این است که نتایج طبقه‌بندی‌کننده‌های موجود را با هم ترکیب کنیم (هم‌جوشی طبقه‌بندی‌کننده‌ها Classifier Fusion) [۹]. هر کدام از طبقه‌بندی‌کننده‌ها بدون در نظر گرفتن این که ممکن است داده‌های ورودی ناقص یا خراب باشند تولید خطا می‌کنند. با این وجود هر کدام از طبقه‌بندی‌کننده‌ها، جنبه‌های گوناگونی از مساله را مد نظر قرار می‌دهند. با فرض این‌که تک تک این طبقه‌بندی‌کننده‌ها، خوب عمل کنند ترکیب آن‌ها خطای کلی طبقه‌بندی را کاهش می‌دهد و نهایتاً منجر به نتایج بهتری می‌شود. به این منظور، در این مقاله از روش هم‌جوشی اطلاعات جهت بهبود کارایی و افزایش یقین به نتایج استفاده شده است.

این مقاله، به صورت زیر سازماندهی شده است. در بخش دوم، به بررسی روش‌های انجام کار پرداخته شده و الگوریتم پیشنهادی شرح داده شده است. بخش سوم، به بررسی و بحث روی نتایج می‌پردازد. در نهایت در بخش آخر، به جمع‌بندی پرداخته شده است.

در این مقاله، روشی جدید مبتنی بر آنالیز اجزای اصلی و هم‌جوشی طبقه‌بندی‌کننده‌ها، برای تشخیص اتوماتیک بیماری هپاتیت پیشنهاد شده است. شکل ۱، ساختار کلی الگوریتم پیشنهادی را نشان می‌دهد.



شکل ۱. ساختار کلی الگوریتم پیشنهادی

مواد و روش‌ها

مجموعه داده‌ها. مجموعه داده‌های بیماری هپاتیت، از انباره داده‌های پایگاه داده‌ی UCI [۶] گرفته شده است. هدف این مجموعه داده‌ها، پیش‌بینی وجود یا عدم وجود این بیماری

با توجه به تست‌های پزشکی انجام شده روی بیماران مختلف است. این مجموعه داده‌ها، دارای ۱۵۵ نمونه است که ۳۲ مورد از نمونه‌ها متعلق به کلاس "die" و بقیه ۱۲۳ مورد مربوط به کلاس "live" می‌باشد. هر نمونه در کلاس، دارای ۱۹ ویژگی می‌باشد. تمام این ۱۹ ویژگی در جدول ۱ لیست شده‌اند. در این مجموعه داده‌ها، ویژگی‌هایی مفقودی وجود دارد که تعداد آن‌ها به ازای هر ویژگی در جدول ۱ نشان داده شده است.

جدول ۱. مشخصات ویژگی‌های مربوط به مجموعه داده‌ها

شماره ویژگی	نام ویژگی	دامنه ویژگی	تعداد مقادیر مفقودی
۱	Age	10, 20,30,40,50,60, 70,80	0
۲	Sex	Male, Female	0
۳	Steroid	Yes, No	1
۴	Antiviral	Yes, No	0
۵	Fatigue	Yes, No	1
۶	Malaise	Yes, No	1
۷	Anorexia	Yes, No	1
۸	Liver big	Yes, No	10
۹	Liver Firm	Yes, No	11
۱۰	Spleen palpable	Yes, No	5
۱۱	Spiders	Yes, No	5
۱۲	Ascites	Yes, No	5
۱۳	Varices	Yes, No	5
۱۴	Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00	6
۱۵	Alk phosphate	33, 80, 120, 160, 200, 250	29
۱۶	SGOT	13,100,200,300, 400,500	4
۱۷	ALBUMIN	2.1, 3.0, 3.8, 4.5, 5.0, 6.0	16
۱۸	PROTIME	10, 20, 30, 40, 50, 60, 70,80, 90	67
۱۹	HISTOLOGY	Yes, No	0

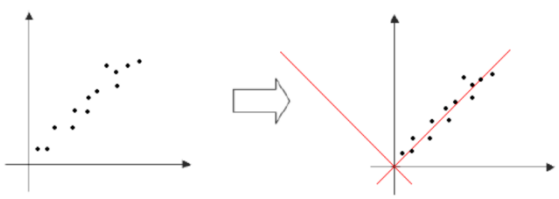
پیش‌پردازش داده‌ها. از آن‌جایی که در این مجموعه داده‌ها، ویژگی‌هایی با مقادیر مفقودی وجود دارد. در این بخش، به عنوان یک فاز پیش‌پردازش، مقادیر این ویژگی‌های مفقودی با استفاده از روشی مبتنی بر شباهت [۱۰] پیش‌بینی شده است. در این روش، به ازای هر نمونه‌ای که دارای مقادیر مفقودی است با استفاده از فاصله اقلیدسی نزدیک‌ترین نمونه به آن پیدا شده و مقادیر نامعلوم این نمونه، با استفاده از نزدیک‌ترین نمونه پیدا شده مشخص می‌گردد.

نرمال‌سازی داده‌ها. از آن‌جایی که ویژگی‌های استخراج‌شده، دارای واحدهای مختلفی هستند جهت یکسان‌سازی ارزش آن‌ها مقادیر هر ویژگی، به فاصله بین ۰ تا ۱ برده شده است. در این فرایند نرمال‌سازی، اگر کوچک‌ترین عدد مربوط به یک ویژگی با Min و بزرگ‌ترین عدد با Max نشان داده شود عدد نرمال شده‌ی X به صورت زیر به دست می‌آید:

$$y = \frac{x - \min}{\max - \min}$$

کاهش ابعاد داده‌ها با استفاده از آنالیز اجزای اصلی.

در سال‌های اخیر، آنالیز اجزای اصلی (PCA)، برای بعضی کاربردهای شناسایی الگو مثل پردازش سیگنال، شناسایی چهره و فشرده‌سازی تصاویر به وفور استفاده شده است. PCA روشی برای کاهش ابعاد داده‌ها می‌باشد. در PCA محورهای جدیدی برای داده‌ها تعریف شده و داده‌ها بر اساس این محورهای مختصات جدید بیان می‌شوند. اولین محور باید در جهتی قرار گیرد که واریانس داده‌ها بیشینه شود و دومین محور باید عمود بر محور اول به گونه‌ای قرار گیرد که واریانس داده‌ها در آن جهت بیشینه شود به همین ترتیب محورهای بعدی بر تمام محورهای قبلی به گونه‌ای قرار می‌گیرند که داده‌ها در آن جهت دارای بیش‌ترین پراکندگی باشند. در شکل ۲، این مفهوم برای داده‌های دو بعدی نشان داده شده است.



شکل ۲. محورهای جدید در جهت پرتراکم ترین نقاط قرار دارند.

الگوریتم کاهش ابعاد با استفاده از PCA، در زیر شرح داده شده است: M یک مجموعه داده t بعدی است. n محور اساسی G1,G2,…,Gn بر هم عمود هستند. در حالت کلی، G1,G2,…,Gn می‌تواند به وسیله n بردار ویژه از ماتریس کوواریانس نمونه‌ها به‌دست 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 نمونه‌ها و L تعداد نمونه‌ها است. با توجه به این مطلب:

$$UG_k=v_kG_k, k\in 1,...,n$$

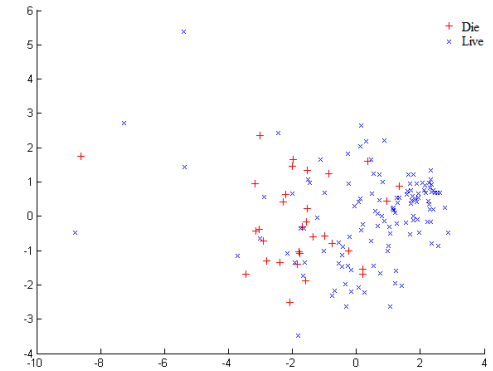
که v_k، k تا بزرگ‌ترین مقدار ویژه U است. n جز اصلی یک بردار $x_k \in M$ به صورت زیر داده شده است:

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

که q، n تا اجزای اصلی x می‌باشند.

نمودار پراکندگی مربوط به دو ویژگی جدید با استفاده از دو جز اصلی از مجموعه داده‌های بیماری هپاتیت در شکل ۳ نشان داده شده است.

مدل‌سازی با استفاده از طبقه‌بندی‌کننده‌های مختلف. بعد از فرایند کاهش تعداد ویژگی‌ها، از طبقه‌بندی‌کننده‌های مختلف جهت مدل‌سازی رابطه بین ورودی‌ها و خروجی‌ها استفاده شده است. به این منظور طبقه‌بندی‌کننده‌های مختلف مورد استفاده قرار گرفته است که در زیر شرح کوتاهی در مورد این طبقه‌بندی‌کننده‌ها داده شده است.



شکل ۳: نمودار پراکندگی مربوط به داده ها با استفاده از دو جز اصلی

شبکه‌های عصبی پس انتشار. شبکه‌های عصبی از تعدادی واحدهای پردازشی کوچکی به نام نرون تشکیل شده است که مجموعه ورودی را به خروجی ربط می‌دهد. یکی از ساده‌ترین و در عین حال کارامدترین چیدمان‌های پیشنهادی برای استفاده در مدل‌سازی عصب‌های واقعی، مدل پرسپترون چندلایه (MLP) می‌باشد که از یک لایه ورودی، یک یا چند لایه پنهان و یک لایه خروجی تشکیل شده است. در این ساختار، تمام نرون‌های یک لایه به تمام نرون‌های لایه بعد متصل‌اند. به منظور آموزش شبکه و اصلاح وزن‌ها تا رسیدن به یک خطای معنادار، روش‌های بسیار زیادی وجود دارد. یکی از مشهورترین این روش‌ها، الگوریتم پس‌انتشار خطا [۱۱] است که در این‌جا از این الگوریتم آموزشی استفاده شده است.

ماشین‌بردار پشتیبان. استفاده از ماشین‌های بردارهای پشتیبان (SVM) در مسائل طبقه‌بندی، رویکرد جدیدی است که در چند ساله اخیر مورد توجه بسیاری قرار گرفته است و از آن در طیف وسیعی از کاربردها از جمله OCR، تشخیص دست‌خط، تشخیص علائم راهنمایی و ... استفاده کرده‌اند. رویکرد SVM به این صورت است که در فاز آموزش، سعی می‌شود که مرز تصمیم‌گیری به گونه‌ای انتخاب گردد که حداقل فاصله آن با هر یک از دسته‌های مورد نظر بیشینه گردد. این نوع انتخاب باعث می‌شود که تصمیم‌گیری ما در عمل، شرایط نویزی را به خوبی تحمل کند و پاسخ‌دهی خوبی داشته باشد. این نحوه انتخاب مرز بر اساس نقاطی به نام بردارهای پشتیبان انجام می‌شود.

فرض کنید که کلاس‌ها به صورت خطی قابل تفکیک شدن هستند از این رو یک صفحه (یا یک خط در حالت دوبعدی) می‌تواند این دو را از هم جدا کند. معادله کلی این صفحه به صورت زیر است:

$$w^Tx+b=0$$

بنابراین $w^Tx+b=+1$ صفحه‌ای (صفحه مثبت) است که بردارهای پشتیبان در کلاس اول هستند و $w^Tx+b=-1$ (صفحه منفی) صفحه‌ی مربوط به بردارهای پشتیبان در کلاس

دیگر می‌باشد. برای نمونه ناشناخته x، اگر معادله $w^T x+b\geq 1$ برقرار باشد نمونه متعلق به اولین کلاس و در صورتی که $w^T x+b\leq -1$ باشد نمونه متعلق به کلاس دوم می‌باشد.

اگر x+ یک نمونه روی صفحه مثبت باشد و x− نزدیک‌ترین نقطه به x+ روی صفحه‌ی منفی باشد پهنای حاشیه M می‌تواند به صورت فاصله‌ی بین x+ و x− بیان شود:

$$M = | x^+ - x^- |$$

هرچند که ثابت می‌شود که M می‌تواند بر حسب W به صورت زیر نیز تعریف شود:

$$M = \frac{2}{\sqrt{w^T w}}$$

از فرمول بالا، حاشیه ماکزیمم M هم‌ارز با کمینه کردن $\sqrt{w^T w}$ است. بنابراین مساله پیدا کردن بهترین طبقه‌بندی‌کننده می‌تواند به صورت یک مساله بهینه‌سازی تصور شود که تابع هدف کمینه کردن حاشیه M است:

$$\min_{w,b} \frac{1}{2} w^T w$$

و شرط‌های محدودکننده به صورت زیر بیان می‌شوند:

$$\begin{cases} w^T x^i + b \geq +1 & \forall i \in C^+ \\ w^T x^i + b \leq -1 & \forall i \in C^- \end{cases}$$

حل این معادله، یک مسئله بهینه‌سازی درجه دوم است که می‌تواند با تبدیل به یک مسئله هم‌ارز حل شود. این فرمول به صورت زیر بیان می‌شود.

$$\sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

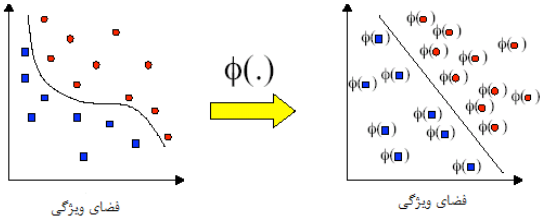
با شرط‌های

$$\sum_{i=1}^N \alpha_i y_i = 0,$$

$$0 \leq \alpha_i \leq C, \quad \forall i$$

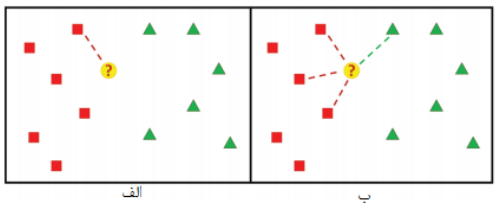
در صورتی که داده‌ها به صورت خطی تفکیک‌پذیر نباشند از یک تابع هسته استفاده می‌شود که داده‌ها را به یک فضای ویژگی دیگر با قابلیت تفکیک خطی می‌برد (شکل ۴).

روش بحث‌شده برای ماشین‌بردار پشتیبان برای مساله طبقه‌بندی دوکلاسه بود. ماشین‌بردار پشتیبان برای مسائل چندکلاسی نیز قابل تعمیم است. معمولاً به این منظور از دو روش "یکی در برابر بقیه" و "یک در برابر یک" استفاده می‌شود [۱۲]. در روش "یکی در برابر بقیه" برای هر کلاس یک ماشین‌بردار پشتیبان ساخته می‌شود که آن کلاس را از بقیه (M-1) کلاس جدا می‌کند. تعداد SVM ها در این روش برابر M است. یک نمونه، به کلاسی که دارای بیش‌ترین مقدار دقت برای جداکنندگی باشد تعلق می‌گیرد. در روش "یک دربرابر یک"، SVM برای هر جفتی از کلاس‌ها اجرا می‌شود. تعداد SVM های مورد استفاده در این روش برابر $M(M-1)/2$ است. برای تشخیص تعلق یک نمونه به یک کلاس، از رای‌گیری استفاده می‌شود.



شکل ۴. استفاده از یک تابع هسته برای بردن داده‌ها به یک فضای ویژگی خطی

نزدیک‌ترین همسایه. ایده پایه‌ای در K نزدیک‌ترین همسایه این است که نمونه‌های مشابه در طبقه‌بندی مشابه با هم قرار می‌گیرند. در K نزدیک‌ترین همسایه، شباهت بین دو نمونه با یک معیار مشخص اندازه‌گیری می‌شود. در این الگوریتم، فاصله‌ی بین نمونه تست با همه‌ی نمونه‌ها اندازه‌گیری می‌شود و K تا از نزدیک‌ترین همسایه‌ها پیدا می‌شوند مثلاً در NN−۱ تمام فاصله‌ها بین نمونه‌ی تست با تمام نمونه‌های آموزشی اندازه‌گیری می‌شود و کلاس نمونه‌ی تست، کلاس مربوط به نزدیک‌ترین نمونه در نظر گرفته می‌شود (شکل ۵).



شکل ۵: (الف) ۱. نزدیکترین همسایه (ب) ۴. نزدیکترین همسایه

معیارهای متفاوتی برای محاسبه‌ی فاصله‌ی بین نمونه‌ها وجود دارد. در این پژوهش، فاصله‌ی اقلیدسی برای اندازه‌گیری معیار شباهت استفاده شده است. فاصله اقلیدسی به صورت زیر تعریف می‌شود.

$$D(i,j)=[\sum_{m=1}^{d}(x_i-x_j)^2]^{\frac{1}{2}}$$

که D(i,j) فاصله‌ی اقلیدسی بین نمونه‌ی i و j، d تعداد ویژگی‌ها، xi موقعیت نمونه تست و xk موقعیت نمونه آموزشی است.

هم‌جوشی طبقه‌بندی‌کننده‌ها. طبقه‌بندی‌کننده‌های مختلف، دارای محدودیت‌های گوناگون هستند و توسعه یک طبقه‌بندی‌کننده‌ی بهتر، کار مشکلی به نظر می‌رسد. یک راه‌کار جهت افزایش کارایی طبقه‌بندی‌کننده‌ها، این است که نتایج طبقه‌بندی‌کننده‌های موجود را با هم ترکیب کنیم. چنین راه‌کار ترکیبی، برای کاهش نایقینی بسیار ارزشمند است. هر کدام از طبقه‌بندی‌کننده‌ها بدون در نظر گرفتن این که ممکن است داده‌های ورودی ناقص یا خراب باشند تولید خطا می‌کنند. با این وجود هر کدام از طبقه‌بندی‌کننده‌ها، جنبه‌های گوناگونی از مساله را مد نظر قرار می‌دهند. با فرض این‌که تک تک این طبقه‌بندی‌کننده‌ها، خوب عمل کنند هم‌جوشی (ترکیب) آن‌ها خطای کلی طبقه‌بندی را کاهش می‌دهد و نهایتاً منجر به نتایج بهتری می‌شود [۹].

در این مقاله، جهت هم‌جوشی طبقه‌بندی‌کننده‌ها، از روش رای‌گیری وزن‌دار استفاده شده است (شکل ۶):

$$y_i=\sum_{j=1}^{L}w_jC_j$$

که C1,…,CL، طبقه‌بندی‌کننده‌های مختلف هستند و W1,…,WL وزن‌های مربوط به هر کدام از طبقه‌بندی‌کننده‌ها می‌باشند.

در این پژوهش، سه طبقه‌بندی‌کننده‌ی نزدیک‌ترین همسایه، شبکه عصبی و ماشین‌بردار پشتیبان به عنوان طبقه‌بندی‌کننده‌های پایه در نظر گرفته شده‌اند. هر طبقه‌بندی‌کننده که دارای دقت بیش‌تری می‌باشد باید تاثیر بیش‌تری در نتیجه نهایی داشته باشد بنابراین دقت طبقه‌بندی‌کننده به عنوان وزن هر طبقه‌بندی‌کننده در نظر گرفته شده است.

نتایج

برای ارزیابی کارایی روش پیشنهادی، آزمایشاتی روی مجموعه داده‌های هپاتیت انجام شده است. این مجموعه داده‌ها از [۶] قابل دانلود می‌باشد. معمولاً برای مقایسه الگوریتم‌های طبقه‌بندی، از روش اعتبارسنجی منقطع k لایه [۱۳] استفاده می‌شود. در این روش، کل داده‌ها، به k فولد (قسمت) تقسیم می‌شود. الگوریتم طبقه‌بندی k بار اجرا می‌شود. در هر بار اجرا یکی از فولدها به عنوان مجموعه تست و اجتماع (k-1) فولد دیگر برای آموزش استفاده می‌شود به عنوان مثال در k=2، در هر بار آزمایش، داده‌ها به صورت تصادفی به دو قسمت مساوی تقسیم می‌شوند. الگوریتم دوبار اجرا می‌شود. در اجرای اول، قسمت اول داده‌ها به عنوان مجموعه آموزشی و قسمت دیگر به عنوان مجموعه تست استفاده می‌شود. در اجرای دوم جای مجموعه‌های آموزش و تست عوض می‌شود. میانگین دقت طبقه‌بندی در این دو اجرا به عنوان دقت الگوریتم در این آزمایش شناخته می‌شود. در این‌جا برای اطمینان از این‌که توزیع‌های متفاوت کلاس‌ها در یک زیرمجموعه، نتایج خیلی متفاوتی را به دنبال ندارد از تعداد ۲، ۵ و ۱۰ فولد استفاده شده است.

معیار کارایی. جهت ارزیابی کارایی طبقه‌بندی‌کننده‌های مختلف استفاده‌شده از ۳ معیار کارایی دقت, صحت، و فراخوانی، استفاده شده است. دقت طبقه‌بندی‌کننده نشان‌دهنده‌ی تعداد نمونه‌های درست به تعداد کل نمونه‌ها است. معیار صحت نشان‌دهنده‌ی این است که از مجموعه داده‌های تست مربوط به نوع A، چه نسبتی درست تشخیص داده شده است. معیار فراخوانی نشان می‌دهد که اگر داده‌ای، نوع A تشخیص داده شد با چه احتمالی نوع A می‌باشد. دقت و فراخوانی به صورت زیر محاسبه می‌شوند:

$$\text{Precision}_i=\frac{TP_i}{TP_i+FP_i}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i}$$

که TPi، FPi، FNi و TNi به ترتیب نشان‌دهنده‌ی تعداد نمونه‌های مثبت درست، مثبت نادرست، منفی نادرست و منفی درست می‌باشد. این متغیرها در جدول ۲ با عنوان ماتریس اغتشاش برای حالت دو کلاس طبقه‌بندی به نام‌های مثبت و منفی توضیح داده شده‌اند. این دو معیار، با استفاده از معیار F با هم ترکیب شده‌اند که به صورت زیر تعریف شده است:

$$F \text{ معیار} = 2 \times \frac{\text{صحت} \times \text{فراخوانی}}{\text{صحت} + \text{فراخوانی}}$$

جدول ۲. مثبت درست، مثبت نادرست، منفی نادرست و منفی درست

		کلاس واقعی	
		مثبت	منفی
کلاس پیش-بینی شده	مثبت	*TP*(مثبت درست)	*FP*(مثبت نادرست)
	منفی	*FN* (منفی نادرست)	*TN*(منفی درست)

نتایج آزمایشات

برای ارزیابی کارایی روش پیشنهادی، آزمایشاتی روی مجموعه داده‌های هپاتیت انجام شده است. دقت طبقه‌بندی روی داده‌های تست برای مجموعه ویژگی‌های کاهش یافته توسط آنالیز اجزای اصلی در جدول ۳ نشان داده شده است. همان‌طور که در این جدول مشاهده می‌شود بالاترین دقت طبقه‌بندی، ۹۶/۳۲٪، با استفاده از ۱۰ فولد به دست آمده است. علاوه بر این، مقادیر صحت، فراخوانی و معیار F در جدول ۴ نشان داده شده است. علت افزایش دقت طبقه‌بندی‌کننده‌ها با افزایش تعداد فولدها، افزایش تعداد نمونه‌های آموزشی و در نتیجه آموزش بهتر طبقه‌بندی‌کننده‌ها بوده است. نتایج طبقه‌بندی با استفاده از ماتریس اغتشاش برای یکی از اجراها در جدول ۵ نشان داده شده است.

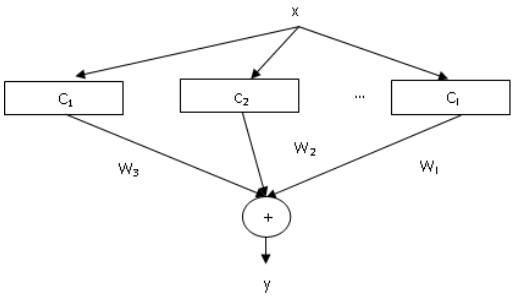
برای مقایسه با کارهای قبلی، جدول ۶ دقت طبقه‌بندی روش پیشنهادی را با روش‌های قبلی نشان می‌دهد. همان‌طور که در این جدول دیده می‌شود روش پیشنهادی توانسته است دقت قابل قبول ۹۶/۳۲٪ نسبت به دیگر مقالات را گزارش دهد.

جدول ۳. دقت طبقه بندی روی داده‌های تست روی تعداد فولدهای گوناگون

دقت طبقه بندی کننده ها(٪)		
۲ فولد	۵ فولد	۱۰ فولد
۷۷,۵	۸۴,۸۳	۹۶,۳۲

مطالعه‌ی مقایسه‌ای. هدف از این آزمایش، مقایسه روش پیشنهادی با سایر پارامترها می‌باشد. در ابتدا تمام داده‌ها در ابتدا به فاصله بین ۰ تا ۱ نرمال شده‌اند. سپس با استفاده از طبقه‌بندی‌کننده‌های نزدیک‌ترین همسایه، شبکه عصبی، ماشین‌بردار پشتیبان، مدل‌سازی انجام گرفته است و نتایج با استفاده از روش ترکیب طبقه‌بندی‌کننده‌ها مقایسه گشته است. نتایج طبقه‌بندی با استفاده از تمام ویژگی‌ها و ۱۰ فولد در جدول ۷ نشان داده شده است. دقت طبقه‌بندی بین ۷۶/۶۶ تا ۸۷/۲۳ با استفاده از تمام ویژگی‌ها به دست آمده است. دقت پایین نتایج، به خاطر وجود ویژگی‌های زائد و غیر مفید می‌باشد که منجر به کاهش دقت طبقه‌بندی‌کننده‌ها شده است. بنابراین در مرحله بعد، تعداد ویژگی‌ها کاهش پیدا کرده است.

بعد از فرایند نرمال‌سازی، PCA برای کاهش ابعاد مورد استفاده قرار گرفت و اولین ۷، ۸، ۹ و ۱۰ جز اصلی از ۱۹ ویژگی اولیه استخراج شد. همان‌طور که در شکل ۶ نشان داده شده است این تعداد از اجزای اصلی بیش‌تر از ۸۲٪ اطلاعات داده‌ها را شامل می‌شوند. تمام ۴ زیرمجموعه‌ی کاهش‌یافته (۷، ۸، ۹ و ۱۰ جز اصلی) وارد طبقه‌بندی‌کننده‌ها شده‌اند.



شکل ۶: رای گیری وزن دار در یک ساختار همجوشی طبقه بندی کننده‌ها

نتایج هم‌جوشی طبقه‌بندی‌کننده‌ها، با استفاده از اولین ۷-۱۰ جز اصلی برای داده‌های نرمال شده در جدول ۸ نشان داده شده است. بازه دقت طبقه‌بندی در فاصله ۸۶,۵۴ تا ۹۶/۳۲ بوده است. در میان آن‌ها، ۱۰ جز اصلی دارای کارایی بهتری بوده است. بیش‌ترین دقت طبقه‌بندی با استفاده از ۱۰ فولد و دقت ۹۶/۳۲ به دست آمده است.

نتایج سه طبقه‌بندی‌کننده‌ی شبکه عصبی، ماشین‌بردار پشتیبان و نزدیک‌ترین همسایه با استفاده از ۱۰ فولد در جدول ۸ آمده است.

در بین این سه طبقه‌بندی‌کننده، ماشین‌بردار پشتیبان بهترین نتایج را ارائه نموده است. مقایسه این نتایج با نتایج جدول قبل، نشان می‌دهد که ترکیب طبقه‌بندی‌کننده‌ها به طور کارایی می‌تواند نتایج تشخیص را بهبود بخشد. هم‌چنین به دلیل این‌که از چند طبقه‌بندی‌کننده استفاده شده اطمینان به نتایج نیز بیش‌تر خواهد بود.

جدول ۴. مقادیر صحت و فراخوانی برای داده های تست روی تعداد فولدهای گوناگون

متریک	تعداد فولدها ۲	۵	۱۰
صحت	۹۷,۳۵	۹۸,۶۴	۹۹,۳۲
فراخوانی	۸۲,۳۵	۸۱,۵۴	۸۶,۷۱
معیار F	۸۹,۲۲	۸۹,۲۸	۹۲,۵۹

جدول ۵. ماتریس اغتشاش برای یکی از اجراها

تعداد فولدها	مقدار پیش بینی شده: بیمار	نرمال	مقدار واقعی
۲	۱	۵۹	نرمال
	۱۳	۴	بیمار
۵	۲	۲۵	نرمال
	۲	۲	بیمار
۱۰	۰	۱۲	نرمال
	۲	۱	بیمار

جدول ۶. مقایسه روش پیشنهادی با دیگر روش‌ها

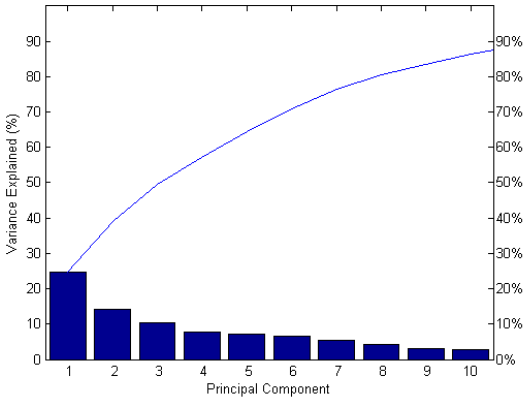
روش	دقت طبقه بندی	مرجع
CSFNN	۹۰,۰۰	[۱۴]
C4.5	۸۳,۶۰	[۱۴]
NB	۸۷,۸۰	[۱۴]
BNND	۹۰,۰۰	[۱۴]
BNNF	۸۸,۷۰	[۱۴]
RBF	۷۹,۰۰	[۱۵]
MLP+BP	۷۷,۴۰	[۱۵]
GRNN[5FC)	۸۰,۰۰	[۱۵]
FS-Fuzzy-AIRS (50-50%)	۸۱,۸۰	[۱۶]
Fuzzy-AIRS with fuzzy res (10FC)	۹۲,۵۰	[۳]
FS-Fuzzy-AIRS(10FC)	۹۴,۱۰	[۱۶]
LDA-ANFIS(10FC)	۹۴,۱۰	[۷]
LFDA-SVM	۹۶,۷۷	[۸]
PCA-LSSVM	۹۵,۰۰	[۷]
GA-SVM	۸۹,۶۰	[۱۷]
CBR-PSO	۹۲,۸۳	[۱۸]
MLNN with levenberg Marquardt	۹۱,۸۷	[۱۹]
MLP-ICA	۹۵,۴۶	[۲۰]
PCA-ANN	۸۹,۶۰	[۲۱]
CART	۸۲,۷۰	[۲۲]
PNN	۹۱,۲۵	[۲۳]
روش پیشنهادی در این مطالعه	۹۶,۳۲	–

جدول ۷. طبقه بندی با استفاده از تمام ویژگی ها

طبقه بندی کننده	دقت طبقه بندی
نزدیکترین همسایه	۷۶,۶۶
شبکه عصبی	۸۰,۵۳
ماشین بردار پشتیبان	۸۴,۴۲
ترکیب طبقه بندی کننده ها	۸۷,۲۳

جدول ۸. نتایج همجوشی طبقه بندی کننده ها با تعداد متفاوت از اجزای اصلی

دقت طبقه بندی	تعداد فولدها	تعداد اجزای اصلی
۸۶,۵۴	۲	۷
۸۹,۲۳	۵	
۹۱,۳۵	۱۰	
۸۸,۶	۲	۸
۹۰,۲۱	۵	
۹۲,۳۵	۱۰	
۸۷,۶۵	۲	۹
۹۱,۷۳	۵	
۹۳,۶۸	۱۰	
۹۰,۵۶	۲	۱۰
۹۳,۴۶	۵	
۹۶,۳۲	۱۰	



شکل ۷. درصد اطلاعات هر جز اصلی

جدول ۸. نتایج طبقه بندی کننده های مورد استفاده

نوع طبقه بندی کننده	دقت طبقه بندی کننده
نزدیکترین همسایه	۹۰,۲۳
شبکه عصبی	۸۵,۶۴
ماشین بردار پشتیبان	۹۴,۳۲

بحث و نتیجه‌گیری

در این مطالعه، سیستمی هوشمند برای تشخیص بیماری هپاتیت با استفاده از همجوشی اطلاعات ارائه شد. بعد از نرمال‌سازی داده‌ها، با استفاده از تکنیک آنالیز اجزای اصلی، تعداد ویژگی‌ها به ۱۰ کاهش پیدا کرد سپس با استفاده از سه طبقه‌بندی‌کننده و همجوشی اطلاعات آن‌ها، طبقه‌بندی انجام شد. نتایج آزمایشات نشان می‌دهد که روش ارائه‌شده کارایی قابل قبولی نسبت به روش‌های قبلی دارد. نرخ دقت طبقه‌بندی ۹۶/۳۲ با استفاده از روش پیشنهادی به دست آمده است بنابراین سیستم پیشنهادی می‌تواند جهت تشخیص نهایی پزشکان بسیار مفید واقع شود.

روش پیشنهادی نتایج چند طبقه‌بندی‌کننده را جهت بهبود کارایی و افزایش دقت به نتایج ترکیب نمود. برای ترکیب طبقه‌بندی‌کننده‌ها در این‌جا از روش رای‌گیری استفاده شد. در اینجا فرض شد که تمام طبقه‌بندی‌کننده‌ها در هر فضای ویژگی ممکن است دارای ضعف باشند. این فرض باعث خواهد شد تا با این‌که یک طبقه‌بندی‌کننده به صورت درستی طبقه‌بندی را انجام داده است اما طبقه‌بندی‌کننده‌های غالب دیگر نتیجه نهایی را به نفع خود تغییر دهند. به این منظور می‌توان قبل از رای‌گیری بررسی کرد که یک طبقه‌بندی‌کننده در چه فضای ویژگی دارای ضعف و قوت است و فقط در فضاهایی که دارای ضعف می‌باشد رای دیگر طبقه‌بندی‌کننده‌ها را دخالت داد. هم‌چنین می‌توان کارایی سیستم تشخیص را با استفاده از روش‌های دیگر ترکیب طبقه‌بندی‌کننده‌ها، بهبود داد.

تشکر و قدردانی

این مطالعه با حمایت مالی دانشگاه کاشان به انجام رسیده است لذا نویسندگان بر خود لازم می‌دانند تا از مدیریت محترم دانشگاه جهت حمایت‌های مالی تشکر و قدردانی نمایند.

منابع

[1] Alter MJ. Epidemiology of hepatitis B in Europe and worldwide. J Hepatol 2003; 39: 64-69.

[2] Pan CQ, Zhang JX. Natural history and clinical consequences of hepatitis B virus infection. Int J Med Sci 2005; 2: 36-40.

[3] Polat K, Güneş S. Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation. Digit Signal Process 2006; 16: 889-901.

[4] Çalişir D, Dogantekin E. A new intelligent hepatitis diagnosis system: PCA–LSSVM. Expert Syst Appl 2011; 38: 10705-10708.

[5] Polat K, Güneş S. Prediction of hepatitis disease based on principal component analysis and artificial immune recognition system. Appl Math Comput 2007; 189: 1282-1291.

[6] Bache K, Lichman M. UCI machine learning repository. Cent Mach Learn Intell Syst 2013; 4: 4. URL http://archive.ics.uci.edu/ml.

[7] Dogantekin E, Dogantekin A, Avci D. Automatic hepatitis diagnosis system based on linear discriminant analysis and adaptive network based on fuzzy inference system. Expert Syst Appl 2009; 36: 11282-11286.

[8] Chen HL, Liu DY, Yang B, Liu J, Wang G. A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis. Expert Syst Appl 2011; 38: 11796-11803.

[9] Kuncheva L. Combining Pattern Classifiers. STUDFUZZ. John Wiley & Sons; 2004.

[10] Hastie T, Tibshirani R, Sherlock G, Eisen M, Brown P, Botstein D. Imputing missing data for gene expression arrays. Stanford Univ Stat Dep Tech 1999.

[11] Hecht-Nielsen R, editor. Theory of the backpropagation neural network. Neural Networks, 1989 IJCNN, International Joint Conference on; 1989: IEEE.

[12] Suykens JA, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett 1999; 9: 293-300.

[13] Bishop CM, Nasrabadi NM. Pattern recognition and machine learning. Springer 2006.

[14] Ozyilmaz L, Yildirim T. Artificial neural networks for diagnosis of hepatitis disease. Neural Network 2003.

[15] Duch W, Adamczak R, Grabczewski K. Optimization of logical rules derived by neural procedures. Neural Network 1999.

[16] Polat K, Güneş S. A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS. Comput Methods Programs Biomed 2007; 88: 164-174.

[17] Tan K, Teoh E, Yu Q, Goh K. A hybrid evolutionary algorithm for attribute selection in data mining. Expert Syst Appl 2009; 36: 8616-8630.

[18] Neshat M, Sargolzaei M, Nadjaran Toosi A, Masoumi A. Hepatitis disease diagnosis using hybrid case based reasoning and particle swarm optimization. ISRN Artific Intell 2012.

[19] Bascil MS, Temurtas F. A study on hepatitis disease diagnosis using multilayer neural network with Levenberg Marquardt training algorithm. J Med Syst 2011; 35: 433-436.

[20] Rezaee Kh, Rasegh Ghezelbash M, Chagha Ghasemi N, Haddania J. An intelligent diagnostic system for detection of hepatitis usi, ng multi-layer perceptron and colonial competitive algorithm. J Math Com Sci 2012; 4: 237-245.

[21] Jilani TA, Yasin H, Yasin MM. Pca-ann for classification of hepatitis c patients. Intern J Comp Appl 2011; 14: 1-6.

[22] Sathyadevi G. Application of CART algorithm in hepatitis disease diagnosis. Rec Trend Inform Technol (ICRTIT) 2011.

[23] Bascil MS, Oztekin H. A study on hepatitis disease diagnosis using probabilistic neural network. J Med Syst 2012; 36: 1603-1606.

A new intelligent hepatitis diagnosis using principal component analysis and classifiers fusion

Seyed jalaleddin Mousavirad (Ph.D)*, Hossein Ebrahimpour-Komleh2 (Ph.D)
Dept. of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Kashan, Kashan, Iran

(Received: 15 Mar 2014; Accepted: 30 Aug 2014)

Introduction: In recent years, hepatitis diseases have become prevalent in the world. The correct diagnosis of hepatitis disease is not a straight task. The goal of this paper is to introduce a new intelligent system for automatic hepatitis diagnosis based on machine learning approaches.

Materials and Methods: the proposed approach consists of three stages, namely dimension reduction, classification, and fusion of classifiers. The hepatitis disease features were obtained from UCI machine learning repository. First, features have been normalized. Then, the number of these features is reduced to 10 from 19 by principal component analysis. In the next step, the reduced features are fed to three classifiers. Finally, a classifiers fusion to improve the efficiency and more reliable results using majority voting is presented.

***Results*:** the proposed approach obtained a classification accuracy of 96.32 via 10 fold cross validation.

Conclusion: according to the results, the proposed system can be used as an intelligent partner for the final hepatitis diagnosis by physician.

Keywords: Intelligent diagnosis, Hepatitis, Disease, Machine learning, Classifiers fusion, Principal component analysis.

* Corresponding author. Fax: +98 31 55912424; Tel +98 09354626334
jalalmoosavirad@gmail.com

How to cite this article:
mousavirad S, Ebrahimpour Komleh H. A new intelligent hepatitis diagnosis using principal component analysis and classifiers fusion. koomesh. 2015; 16 (2) :149-158
URL http://koomeshjournal.semums.ac.ir/browse.php?a_code=A-10-2237-2&slc_lang=fa&sid=1