مجله علمی پژوهشی دانشگاه علوم پزشکی شهید صدوقی یزد
دوره ۲۵، شماره ۴، تیر ۱۳۹۶
صفحه: ۳۱۰-۳۰۰

Journal of Shahid Sadoughi University of Medical Sciences
Vol. 25, No.4, Jul 2017
Pages: 300-310

مطالعات درخت تصمیم در برآورد ریسک ابتلا به سرطان سینه با استفاده از چند شکلی‌های تک نوکلوئیدی

فریدا سیدمیر[۱]، کمال میرزایی[۲]، مرتضی بیطرف ثانی*[۳]

چکیده

مقدمه: درختان تصمیم از ابزارهای داده‌کاوی برای جمع‌آوری، پیش‌بینی دقیق و غربال کردن اطلاعات از حجم عظیم داده‌هاست که کاربرد گسترده‌ای در زمینه زیست‌شناسی محاسباتی و بیوانفورماتیک پیدا کرده است. در بیوانفورماتیک می‌توان پیش‌بینی‌هایی بر روی بیماری‌هایی از جمله سرطان سینه داشت. استفاده از داده‌های ژنومی از جمله چند شکلی‌های تک نوکلئوتیدی در پیش‌بینی ریسک ابتلا به بیماری‌های چند عامله از اهمیت خاصی برخوردار است.

روش بررسی:با مطالعه تحلیلی آینده نگر، احتمال ابتلا به سرطان سینه با استفاده از SNP های مرتبط با فرمول $x_j = f_o * \sum_{i=1}^{m} \ln(OR_i) \times SNP_{ij}$ و درختان تصمیم محاسبه شد. هفت SNP با نسبت‌های مختلف بخت مرتبط با سرطان سینه در نظر گرفته و کدنویسی و طراحی درخت تصمیم مدل C4.5، با زبان برنامه‌نویسی Csharp2013 انجام شد.

نتایج: با روش کدنویسی در دو سناریو با افزایش درصد آموزش از ۶۶/۶۶ به ۸۶/۴۲، خطا از ۵۵/۵۶ به ۹/۰۹ کاهش یافت. همچنین با اجرای نرم‌افزار WEKA در سه سناریو با تعداد مجموعه داده‌های مختلف، تعداد مجموعه آموزش مختلف و آزمایش مختلف با افزایش تعداد رکوردها از ۸۱ به ۲۱۸۷، میزان خطا از ۴۸/۱۵ به ۱۳/۴۶ کاهش یافت. همچنین در اکثر سناریوها درصد شیوع بیماری در میزان خطا در کد و WEKA تأثیری نداشت.

نتیجه‌گیری: نتایج نشان می‌دهد با افزایش میزان آموزش، خطای درخت تصمیم کاهش و درنتیجه دقت پیش‌بینی ریسک ابتلا به سرطان سینه با استفاده از درخت تصمیم افزایش می‌یابد. با افزایش رکوردهای مجموعه آموزش و همچنین افزایش تعداد ویژگی SNP در درخت تصمیم، دقت پیش‌بینی افزایش و خطا کاهش می‌یابد.

واژه‌های کلیدی: درخت تصمیم، سرطان سینه، چندشکلی تک نوکلئوتیدی

۱- گروه مهندسی کامپیوتر، واحد یزد، دانشگاه آزاد اسلامی، یزد
۲- عضو هیات علمی،گروه مهندسی کامپیوتر، واحد میبد، دانشگاه آزاد اسلامی، میبد
۳- مدرس، دانشگاه جامع علمی کاربردی یزد (مرکز تحقیقات و آموزش کشاورزی و منابع طبیعی)
***(نویسنده مسئول)؛ تلفن. ۰۹۱۳۳۵۵۰۰۶۰ ، پست الکترونیکی:** bitaraf@sau.ac.ir
تاریخ دریافت: ۱۳۹۴/۱۰/۸ تاریخ پذیرش: ۱۳۹۵/۴/۳۱

مقدمه

در دنیای کامپیوتری امروز پایگاه‌های داده بخش عظیمی از اطلاعات هستند که به علت در دسترس بودن و فراوانی اطلاعات، داده‌کاوی دارای اهمیت بسیاری شده است. داده‌کاوی یک علم، هنر و تکنولوژی بررسی داده‌های بزرگ و پیچیده است که به کشف الگوهای مفید منجر می‌شود (۱). تکنیک‌های داده‌کاوی (Data Mining) که شاخه‌ای از هوش مصنوعی از سال ۱۹۶۰ محسوب می‌شوند امروزه مورد استفاده قرار می‌گیرند (۲). یکی از تکنیک‌های داده‌کاوی درختان تصمیم (Decision Tree) است که به علت فواید جمع‌آوری انواع مختلف داده‌ها با پیش‌بینی دقیق و فهم آسان، کاربردهای گسترده‌ای در زمینه‌های امور مالی، بازاریابی، آموزش و پرورش، مهندسی، پزشکی، زیست‌شناسی محاسباتی و بخصوص بیوانفورماتیک پیداکرده‌اند (۳،۱). داده‌کاوی در پزشکی باعث افزایش دقت تشخیص، کاهش هزینه و کاهش منابع انسانی می‌شود (۴،۳). همچنین از درختان تصمیم برای پیش‌بینی احتمال بیماری با استفاده از چندشکلی‌های تک نوکلییوتیدی در انسان می‌توان استفاده کرد که نقش مهمی در بیوانفورماتیک دارند (۳). توالی‌یابی ژنوم بشر امکان شناسایی اطلاعات بیش از سه میلیون تک‌نوکلئوتید پلی‌مورفیسم (SNP: Single Nucleotide Polymorphism) در سرتاسر ژنوم برای هر فرد را فراهم نموده که از آن می‌توان برای مطالعات جامع ژنومی (GWAS: Genome-wide Association Study) استفاده کرد. با توجه به حجم اطلاعات و پایگاه داده‌های ژنومی، بیوانفورماتیک نقش مهمی در بررسی پیچیدگی‌های زمینه‌های ژنتیکی بیماری‌های شایع و سرطان‌های انسانی ازجمله سرطان سینه ایفا می‌کند (۵). سرطان سینه (Breast Cncer)یک نوع سرطان نشات گرفته از بافت سینه است که به دو نوع خوش‌خیم و بدخیم طبقه‌بندی می‌شوند (۸،۷،۶). سرطان سینه شایع‌ترین سرطان بعد از سرطان پوست و دومین سرطان در مرگ‌ومیر زنان بعد از سرطان ریه است (۹،۷،۶). اگرچه دانشمندان فاکتورهایی چون سن، عوامل ژنی (ژن سرطان پستان BRACA1، BRCA2)، سابقه فامیلی، فرزند نداشتن یا فرزندآوری بعد از سن ۳۵ سال، اضافه‌وزن، قرار گرفتن مکرر در برابر اشعه ایکس، دوره‌های قاعدگی، الکل و سیگار را در افزایش بروز سرطان سینه مؤثر می‌دانند اما هنوز فاکتورهای زیادی را در تبدیل سلول‌ها به سلول‌های سرطانی شناسایی نکردند (۱۰،۸،۷،۶). تشخیص خودکار سرطان سینه یک مشکل پزشکی در دنیای واقعی است و تشخیص این بیماری در مراحل اولیه آن کلیدی برای درمان است (۸،۳). درخت تصمیم‌گیری یک روش قدرتمند برای طبقه‌بندی و پیش‌بینی در مشکل تشخیص سرطان سینه را فراهم می‌کند. در تحقیقات به عمل آمده در سال ۲۰۰۴ درخت تصمیم با دو SNP برای پیش‌بینی ریسک ابتلا به سرطان سینه طراحی شد (۱۱). بررسی‌ها در سال ۲۰۰۸ نشان داده است که اگر درختان تصمیم با داده‌های با کیفیت بالا آموزش داده شود، پیش‌بینی‌های دقیقی در حوزه بیوانفورماتیک به وجود می‌آورند (۳). اگرچه هیچ‌کدام از پیشگویی‌ها صد در صد نیست اما در سال ۲۰۰۹ محققان عملکرد درخت تصمیم در پیش‌بینی سرطان سینه را در دو گروه یک نفره و نوزده نفره از ناظران بررسی کردند و مشاهده کردند که درختان تصمیم بیشتر به دانش افراد خبره بستگی دارد تا به داده واقعی بنابراین در گروه اول با یک ناظر به نتیجه بهتری از پیش‌بینی رسیدند (۱۲). در سال ۲۰۱۴ محققان در تشخیص مارکر مرتبط با سرطان سینه از بین هزاران مارکر ژنتیکی کاندید از روش‌های ترکیبی درخت تصمیم و بهینه‌سازی ازدحام ذرات (PSODT: Particle Swarm Optimization Decision Tree) بهره بردند که باعث کاهش هزینه‌های محاسباتی و خطرات ناشی از عمل جراحی و افزایش دقت پیش‌بینی گردید (۱۳). پژوهشگران در سال ۲۰۱۶ از الگوریتم بهینه‌سازی ازدحام ذرات دودویی با ساختار سلسله مراتبی (BPSOHS: Binary Particle Swam Optimization With Hierachical Structure) بهره بردند بدین ترتیب که با دو بیت، چهار حالت ژنوتیپ صفر، ژنوتیپ یک با هموزیگوت اصلی، ژنوتیپ دو با هموزیگوت فرعی و ژنوتیپ سه با هتروزیکوت را مد نظر قرار دادند و نسبت بخت را محاسبه

کردند که مزایای قابل‌توجهی در شناسایی گروه شاهد و کنترل داشت (۱۰). همچنین تحقیق دیگری در سال ۲۰۱۶ از الگوریتم تشخیص داده‌های خارج از محدوده (ODA:Outler Detection Algorithm) از مجموعه داده وینکاسین با تعداد نه صفت و دو کلاس و با روش ترکیبی J48 و ODA برای تشخیص سرطان پستان در دو نوع خوش‌خیم و بدخیم صورت گرفت (۹). هدف از این تحقیق تعیین احتمال ابتلا به سرطان سینه با استفاده از مارکرهای ژنتیکی SNP ها برای هر فرد با استفاده از درخت تصمیم است.

روش بررسی

هفت SNP با نسبت‌های مختلف بخت (Odd Ratio) مرتبط با سرطان سینه درنظرگرفته و کدنویسی و طراحی درخت تصمیم با زبان برنامه‌نویسی Csharp 2013 انجام شد. در درخت تصمیم ایجادشده با کدنویسی، چهار SNP مهم مرتبط لحاظ شد. خطای درخت تصمیم در دو حالت کدنویسی و استفاده از نرم‌افزار WEKA ارزیابی و تعداد نمونه آموزش داده‌شده با نمونه‌گیری سیستماتیک استخراج گردید. در مطالعات GWAS با استفاده از سیستم‌های کامپیوتری و نرم‌افزاری، بدون دانش قبلی و موقعیت ژن، امکان تعیین ارتباط SNP ها با بیماری امکان‌پذیر شده است. شبیه‌سازی داده‌های ژنومی با استفاده از پایگاه داده HapMap صورت گرفت. مطابق با جدول (۱) نسبت بخت SNP های مرتبط با سرطان سینه که روی کروموزوم‌های ۶، ۸، ۱۰، ۱۱، ۱۶، ۲، ۵ مستقر هستند، برآورد گردید (۴). سپس احتمال ابتلا به بیماری با استفاده از رابطه (۱) محاسبه و سپس برای هر دامنه ریسک ابتلا، یک کلاس تعریف گردید.

$$X_j = fo \times \sum_{i=1}^{m} \ln(ORi) \times SNPij \quad (۱)$$

X_j: احتمال ابتلا به بیماری، F_O: درصد شیوع بیماری، OR_i: نسبت بخت، SNP_{ij}: ژنوتیپ SNP موردنظر

جدول ۱: پارامترهای شبیه‌سازی چندشکلی‌های تک نوکلئوتیدی مرتبط با سرطان سینه (۴)

کروموزوم	SNP (rs number)	نسبت بخت هتروزیگوت	نسبت بخت هموزیگوت
۲	rs13387042	۱/۲۰	۱/۲۰
۵	rs889312	۱/۳۷	۱/۷۴
۶	rs2180341	۱/۴۵	۱/۴۷
۸	rs13281615	۱/۱۹	۱/۴۴
۱۰	rs2981582	۲/۰۰	۲/۲۳
۱۱	rs3817198	۱/۱۵	۱/۴۵
۱۶	rs3803662	۱/۱۲	۱/۱۹

احتمال ابتلا به بیماری به سه کلاس A (درصد بالا)،B (درصد متوسط) و C(درصد پایین) طبقه‌بندی می‌شود. در این درخت تصمیم‌گیری ابتدا طبق فرمول ذکر شده در روش اول احتمال ابتلا به بیماری محاسبه و سپس فاصله اعداد از روش زیر محاسبه شد و این فاصله به سه قسمت مساوی طبق رابطه (۲) تقسیم شد.

فاصله اعداد = بیشترین عدد- کمترین عدد (۲)

یک سوم محدوده اول به کلاس C، یک سوم محدوده میانی به کلاس B و یک سوم محدوده آخر به کلاس A اختصاص داده شد. اعداد به کلاس‌های A و B و C تبدیل شد. محدوده کلاس‌های سه سناریو در جدول (۲) نشان داده‌شده است.

جدول ۲: محدوده کلاس‌ها در سه سناریوی موجود

نام سناریو	کلاس	شیوع 0/05	شیوع 0/1	شیوع 0/15	شیوع 0/2
اول	C	۰-۰/۰۷	۰-۰/۱۴	۰-۰/۲۱	۰-۰/۲۸
	B	۰/۰۷-۰/۱۴	۰/۱۴-۰/۲۸	۰/۲۱-۰/۴۲	۰/۲۸-۰/۵۶
	A	۰/۱۴-۰/۲۱	۰/۲۸-۰/۴۲	۰/۴۲-۰/۶۳	۰/۵۶-۰/۸۶
دوم	C	۰-۰/۰۷	۰-۰/۱۴	۰-۰/۲۱	۰-۰/۲۸
	B	۰/۰۷-۰/۱۴	۰/۱۴-۰/۲۸	۰/۲۱-۰/۴۲	۰/۲۸-۰/۵۶
	A	۰/۱۴-۰/۲۱	۰/۲۸-۰/۴۲	/۴۲-۰/۶۳	۰/۵۶-۰/۸۶
سوم	C	۰-۰/۰۹	۰-۰/۱۹	۰-۰/۲۸	۰-۰/۳۸
	B	۰/۰۹-۰/۱۹	۰/۱۹-۰/۳۸	۰/۲۸-۰/۵۷	۰/۳۸-۰/۷۶
	A	۰/۱۹-۰/۲۸	۰/۳۸-۰/۵۷	۰/۵۷-۰/۸۵	۰/۷۶-۱/۱۳

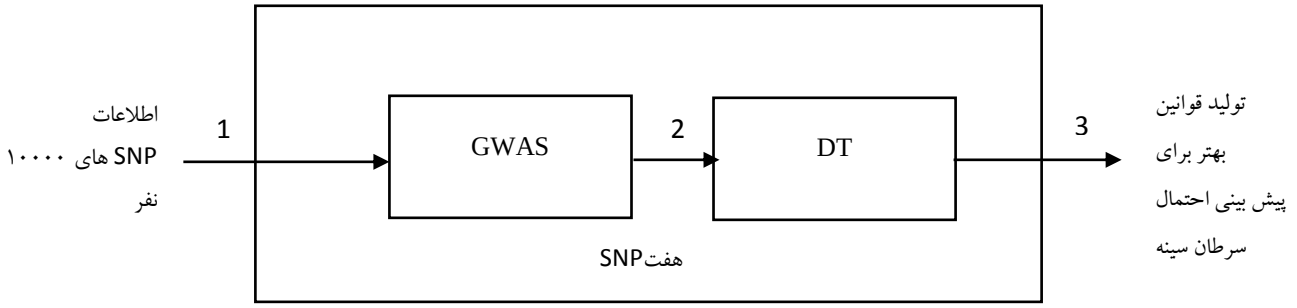
درنهایت مجموعه آموزش (Training Data) و آزمایش (Test Data) طبق سناریوهایی در جدول (۳) ایجاد و به ترتیب در نرم‌افزار Csharp و WEKA پیاده‌سازی و اجرا گردید تا خطای مدل محاسبه و دقت درخت تصمیم پیاده‌سازی شده و اجرا شده برآورد و مقایسه گردد. مطابق با جدول (۳) سه سناریو مورد بررسی قرار گرفت. در هر سناریو به روش سیستماتیک جداول آموزش و آزمایش از کل رکوردها ایجاد شدند. سپس احتمال ابتلا به بیماری به سه کلاس A، B و C مطابق با جدول (۲) محاسبه گردید.

جدول ۳: مشخصات سه سناریوی بکار رفته

سناریو	کل رکوردها	تعداد رکورد جدول آموزش	تعداد رکورد جدول آزمایش	روش انتخاب رکوردها در جدول آموزش به صورت سیستماتیک	روش انتخاب رکوردها در جدول آزمایش به صورت سیستماتیک	کوروموزوم های مورد بررسی
اول	۸۱	۵۴	۲۷	دو رکورد به صورت دو در میان (رکورد ۱ و۲و۴و۵و۷و۸و...و۷۹و۸۰)	یک رکورد به صورت دو در میان(رکورد ۳و۶و۹...و۸۱)	۱۱و۱۰،۶،۵
دوم	۸۱	۷۰	۱۱	هفت رکورد به صورت هفت در میان (رکوردهای ۲و ۳و ۴و ۵و ۶و ۷و ۸و ۱۰ و۱ ۱و ... و ۷۱ و ۷۲ و ۷۴ و ۷۵ و ۷۶ و ۷۷ و ۷۸ و ۷۹ و ۸۰)	یک رکورد به صورت هفت در میان (رکورد۱و۹و۱۷و...و۷۳و۸۱)	۱۱و۱۰،۶،۵
سوم	۲۱۸۷	۱۵۳۳	۶۵۴	هفت رکورد به صورت هفت در میان (رکوردهای ۱و ۲ و ۳ و ۴ و ۵ و ۶ و ۷ و ۱۱ و ... و ۲۱۸۶ و ۲۱۸۷)	سه رکورد به صورت هفت در میان(رکورد۸و۹و۱۰و۱۸و۱۹و۲۰و... و۲۱۷۰و۲۱۷۸و۲۱۷۹و۲۱۸۰	۱۱،۱۰،۸،۶، ۵،۲و ۱۶

درخت تصمیم با مدل C4.5، معیار نسبت بهره (Gain Ratio)، روش هرس بعدی با مدل جایگزینی زیردرخت طراحی شد. از روش بازگشتی (Recursive) در طراحی الگوریتم استفاده نشد. روش کنترل داده‌های گمشده (Lost Data) با برچسب گذاری ERROR، جزء خطای درخت محسوب گردید. برای چهار SNP موثرتر در بین هفت SNP در ارتباط با سرطان سینه روی کروموزوم‌های ۵، ۶، ۱۰، ۱۱، سه ژنوتیپ ۰، ۱، ۲ با دو سناریو طراحی شد؛ بنابراین ۸۱ رکورد متفاوت از ژنوتیپ‌های صفر، یک و دو برای چهار کروموزوم مذکور به وجود آمد یعنی با سه ژنوتیپ (m) و چهار SNP (n)

طبق (m^n) محاسبه شد. مجموعه‌های آموزش و آزمایش برای ترسیم درخت تصمیم به صورت سیستماتیک انتخاب شد. داده‌های تولید شده در محیط Excel و Notpad با فرمت csv و arff وارد محیط WEKA با مدل درخت J48 قرار گرفت. در این حالت نیز فاز سیستماتیک برای سه سناریو مورد بررسی قرار گرفت. دو سناریوی اول همانند کد پیاده‌سازی شده در مرحله قبل و دو سناریو با هفت کروموزوم ۲،۵،۶،۸،۱۰،۱۱،۱۶ و ۲۱۸۷ رکورد اجرا گردید و افزایش رکوردها از ۸۱ به ۲۱۸۷ نمونه بررسی شد. مطابق با شکل (۱) درخت تصمیم با GWAS ترکیب و قوانین بهتری ارائه می‌دهد. در مرحله ۱ (مرحله پیش‌پردازش) با شبیه‌سازی توسط نرم‌افزار GWASimulator، اطلاعات SNP های ۱۰۰۰۰ نفر دریافت و پس از عمل پردازش، در مرحله ۲ هفت SNP روی کروموزوم‌های ۲، ۵، ۶، ۸، ۱۰، ۱۱ و ۱۶ در ارتباط با سرطان سینه به درخت تصمیم تحویل داده می‌شود. درنهایت در مرحله ۳ درخت تصمیم قوانین بهتری را برای پیش‌بینی احتمال سرطان سینه ارائه می‌دهد.

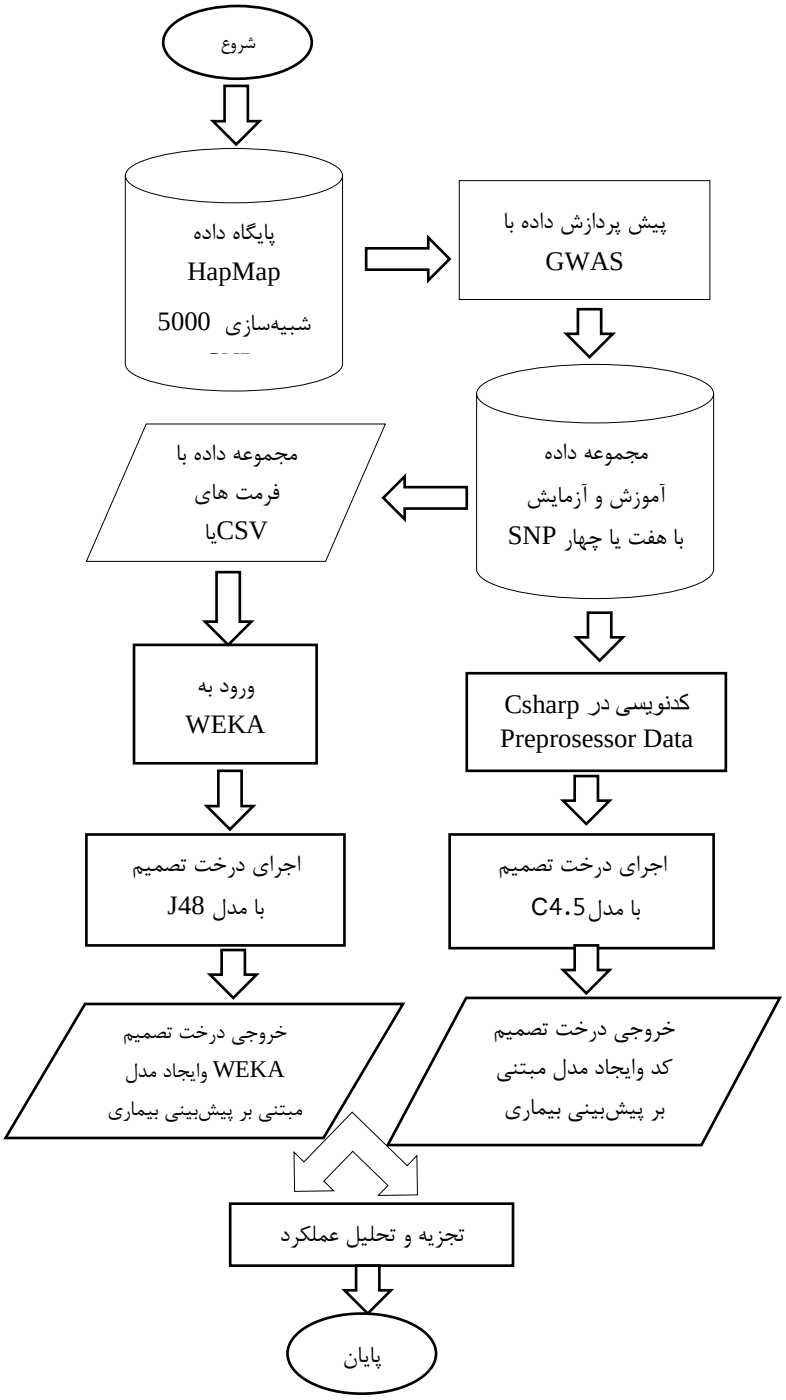


شکل ۱: ترکیب درخت تصمیم و ابزار GWAS برای تولید قوانین بهتر

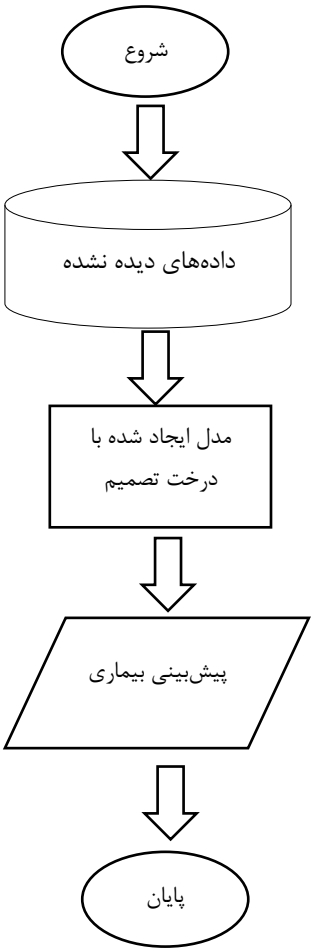
در شکل (۲) نمودار گردشی پیش‌بینی سرطان سینه بر اساس GWAS و DT در دو مرحله داده‌های از قبل معلوم و داده‌های دیده نشده نشان داده‌شده است.

الف) در فاز اول و دوم، مدل GWAS و مدل درخت تصمیم با هم ترکیب می‌شوند. بدین گونه که در فاز اول یعنی در مرحله پیش‌پردازش اطلاعات پنج هزار SNP از ده هزار نفر مربوط به پایگاه داده HapMap وارد GWAS می‌شود. در GWAS با استفاده از نرم‌افزار R روی طیف گسترده‌ای از SNP ها شبیه‌سازی صورت می‌گیرد. افراد در این پایگاه به دو گروه مورد (افراد مبتلا به سرطان سینه) و شاهد (افراد سالم) تقسیم می‌شوند. سپس از بین آن‌ها هفت SNP مؤثر در سرطان سینه شناسایی و از GWAS خارج می‌شود. در فاز دوم چهار SNP با ۸۱ رکورد و هفت SNP با ۲۱۸۷ رکورد به دو قسمت مجموعه آموزش و آزمایش تقسیم و به ترتیب جهت کدنویسی به نرم‌افزار Csharp و اجرا در WEKA فرستاده می‌شود و با پنج سناریو احتمال ابتلا به سرطان سینه محاسبه می‌شود. درنهایت عملکرد مدل پیش‌بینی‌کننده تجزیه و تحلیل می‌گردد.

ب) در فاز سوم داده‌های دیده نشده وارد مدل پیش‌بینی‌کننده می‌شود و احتمال ابتلا به سرطان سینه با دقت مدل پیش‌بینی محاسبه می‌گردد. این مدل پیشنهادی ترکیب درخت تصمیم و GWAS است که می‌تواند بر روی SNP های مؤثر در سرطان سینه بیماری را پیش‌بینی کند بدین ترتیب که طبق معلومات دنباله‌ای با ۵۰۰۰ SNP وارد GWAS شده و پس از اعمال نرم‌افزار R دنباله‌ای با هفت SNP خارج و پس از این معلومات در این پروژه اطلاعات خروجی به درخت تصمیم تحویل داده می‌شود تا پیش‌بینی دقیق‌تری را فراهم کند



الف) فاز اول و دوم



ب) فاز سوم

شکل ۲: نمودار گردشی مدل پیشنهادی پیش‌بینی سرطان سینه

نتایج

در سناریوی اول خطای درخت تصمیم با ۵۴ رکورد آموزش با سه سطح احتمال ابتلا به سرطان سینه به صورت سیستماتیک مورد بررسی قرار گرفت. محاسبه خطاها در تمامی موارد برای چهار درصد شیوع بر روی چهار کوروموزوم در جدول (۴) نشان داده‌شده است.

جدول ۴: نتایج سناریوی بررسی خطای درخت تصمیم با ۵۴ رکورد آموزش با سه سطح احتمال ابتلا به سرطان سینه به صورت سیستماتیک (سناریوی اول)

ردیف	درصد شیوع بیماری	خطا در کد	خطا در وکا
۱	۰/۰۵	۵۵/۵۶	۴۸/۱۵
۲	۰/۱	۵۵/۵۶	۴۸/۱۵
۳	۰/۱۵	۵۵/۵۶	۴۸/۱۵
۴	۰/۲	۵۵/۵۶	۴۸/۱۵

در سناریوی دوم خطای درخت تصمیم با ۷۰ رکورد آموزش با سه سطح احتمال ابتلا به سرطان سینه به صورت سیستماتیک مورد بررسی قرار گرفت. نمونه‌ای از محاسبه خطا در کد نوشته‌شده و WEKA برای چهار درصد شیوع در جدول (۵) نشان داده‌شده است.

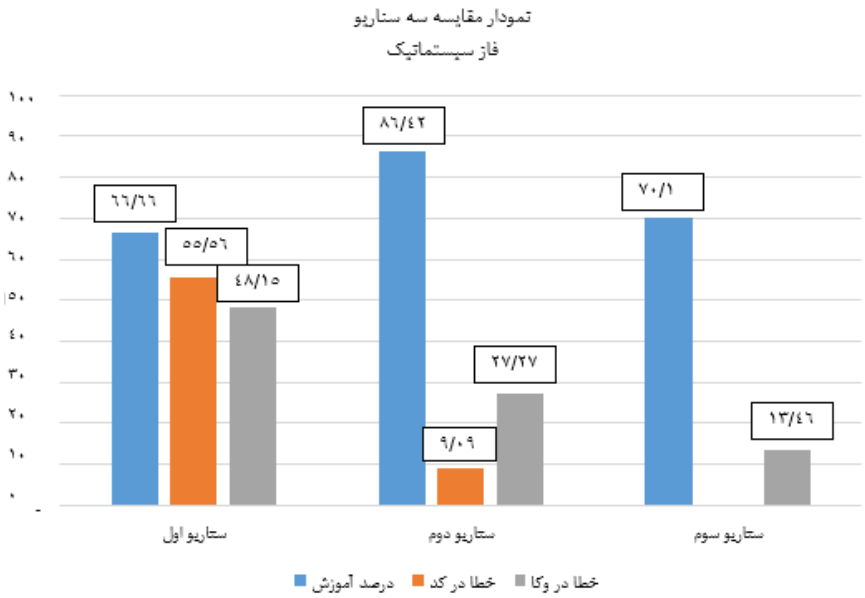
جدول ۵: نتایج سناریوی بررسی خطای درخت تصمیم با ۷۰ رکورد آموزش با سه سطح احتمال ابتلا به سرطان سینه به صورت سیستماتیک (سناریوی دوم)

ردیف	درصد شیوع بیماری	خطا درکد	خطا در وکا
۱	۰/۰۵	۹/۰۹	۲۷/۲۷
۲	۰/۱	۹/۰۹	۲۷/۲۷
۳	۰/۱۵	۹/۰۹	۲۷/۲۷
۴	۰/۲	۹/۰۹	۲۷/۲۷

در سناریوی سوم خطای درخت تصمیم با ۱۵۳۳ رکورد آموزش با سه سطح احتمال ابتلا به سرطان سینه به صورت سیستماتیک مورد بررسی قرار گرفت. محاسبه خطاها در نرم‌افزار WEKA برای چهار درصد شیوع در جدول (۶) نشان داده‌شده است.

جدول ۶: نتایج سناریوی بررسی خطای درخت تصمیم با ۱۵۳۳ رکورد آموزش با سه سطح احتمال ابتلا به سرطان سینه به صورت سیستماتیک (سناریوی سوم)

ردیف	درصد شیوع بیماری	خطا در وکا
۱	۰/۰۵	۱۳/۴۶
۲	۰/۱	۱۳/۴۶
۳	۰/۱۵	۱۳/۴۶
۴	۰/۲	۱۳/۴۶



نمودار ۱: مقایسه میزان خطای درخت تصمیم پیش‌بینی احتمال ابتلا به سرطان سینه بر اساس تعداد SNP و درصد آموزش در سه سناریو به صورت سیستماتیک

در دو حالت کد طراحی شده و نرم‌افزار WEKA خلاصه سناریوها در جدول (۷) و مقایسه سناریوها در نمودار (۱) نشان داده شده است.

جدول ۷: کمترین-بیشترین-میانگین میزان خطای درخت تصمیم پیش‌بینی احتمال ابتلا به سرطان سینه بر اساس تعداد SNP و درصد آموزش در سه سناریو در فاز سیستماتیک

سناریو	تعداد SNP	درصد آموزش	کمترین میزان خطا در کد	کمترین میزان خطا در وکا	بیشترین میزان خطا در کد	بیشترین میزان خطا در وکا	میانگین میزان خطا در کد	میانگین میزان خطا در وکا
اول	۴	۶۶/۶۶	۵۵/۵۶	۴۸/۱۵	۵۵/۵۶	۴۸/۱۵	۵۵/۵۶	۴۸/۱۵
دوم	۴	۸۶/۴۲	۹/۰۹	۲۷/۲۷	۹/۰۹	۲۷/۲۷	۹/۰۹	۲۷/۲۷
سناریو	تعداد SNP	درصد آموزش	کمترین میزان خطا در وکا		بیشترین میزان خطا در وکا		میانگین میزان خطا در وکا	
سوم	۷	۷۰/۱	۱۳/۴۶		۱۳/۴۶		۱۳/۴۶	

خطای درخت تصمیم در دو حالت کدنویسی و استفاده از نرم‌افزار WEKA با اندازه مختلف آموزش به شرح جدول (۸) محاسبه گردید.

جدول ۸: اثر میزان آموزش روی خطای درخت تصمیم پیش‌بینی احتمال ابتلا به سرطان سینه

خطای درخت تصمیم		درصد آموزش	تعداد SNP درخت تصمیم
نرم‌افزار WEKA	کدنویسی		
٪۴۸/۱۴	٪۵۵/۶	٪۶۶	۴
٪۲۷/۱۲	٪۹/۰۹	٪۸۶	۴
٪۱۳/۴۵	-	٪۷۰	۷

بحث

تاکنون گزارشی درخصوص محاسبه ریسک ابتلا به سرطان سینه با استفاده از چندشکلی‌های تک نوکلئوتیدی در ایران با استفاده از درخت تصمیم عنوان نشده است، هرچند برخی گزارش‌ها با استفاده از برخی عوامل محیطی و غیر ژنتیکی و یا شبکه‌های عصبی یا SVM این موضوع را مطرح نموده‌اند (۹،۱۰،۱۳). غالب طرح‌های قبلی عوامل غیر ژنتیک را دخالت داده‌اند و صفات از نوع SNP در این تحقیق برای اولین بار در ایران استفاده شده است. البته در سال ۲۰۰۴ با تعداد دو SNP در درخت تصمیم، پیش‌بینی بیمار یا سالم بودن فرد مورد بررسی قرار گرفت (۱۱) که در این تحقیق با هفت SNP پیش‌بینی صورت گرفته است؛ بنابراین با ترکیب GWAS و درخت تصمیم می‌توان درصد ریسک ابتلا به سرطان سینه را پیش‌بینی نمود.

نتایج در سناریوی اول نشان می‌دهد که درصد خطا برای چهار درصد شیوع بیماری در کد و در نرم‌افزار WEKA یکسان است؛ اما درصد احتمال خطا در وکا پایین‌تر از کد است. به نظر می‌رسد به دلیل اینکه از روش سیستماتیک برای انتخاب مجموعه آموزش و آزمایش انتخاب‌شده است در اکثر سناریوها درصد شیوع بیماری در میزان خطا در کد و WEKA تأثیر ندارد.

نتایج در سناریوی دوم نشان می‌دهد که درصد خطا برای چهار درصد شیوع بیماری در کد و در نرم‌افزار WEKA یکسان است؛ اما درصد احتمال خطا در وکا بالاتر از کد است. البته با یک نمونه نمی‌توان دلیل بهتر یا بدتر بودن عملکرد کد نسبت به WEKA را مشخص نمود. به نظر می‌رسد به دلیل اینکه از روش سیستماتیک برای انتخاب مجموعه آموزش و آزمایش انتخاب شده است درصد شیوع بیماری در میزان خطا در کد و WEKA تأثیر ندارد. با تعداد بیشتر رکوردها در مجموعه آموزش در روش سیستماتیک در این سناریو درصد خطا نسبت به سناریوی اول کاهش یافت. در این نمونه درصد احتمال خطای کد نسبت به WEKA به‌طور چشمگیری کاهش داشت.

سناریوی سوم فقط در نرم‌افزار WEKA اجرا شده است. با توجه به افزایش تعداد ویژگی از چهار به هفت و به دنبال آن افزایش تعداد رکوردها از ۸۱ به ۲۱۸۷ (۱۵۳۳ آموزش، ۶۵۴ آموزش) احتمال خطا به‌طور چشمگیری نسبت به چهار سناریوی قبل کاهش می‌یابد. اگر بتوان نه‌تنها ۳۰ SNP بلکه همه ۳۰۰۰۰۰ SNP را داخل مدل نمود تا با خطای کمتر مواجه شد. احتمال خطا در چهار احتمال شیوع ۱۳/۴۶ است.

با مقایسه این سه سناریو به نظر می‌رسد به دلیل اینکه از روش سیستماتیک برای انتخاب مجموعه آموزش و آزمایش انتخاب شده است در اکثر سناریوها درصد شیوع بیماری در میزان خطا در کد و WEKA تأثیر ندارد. با افزایش درصد آموزش از ۶۶/۶۶ به ۸۶/۴۲ خطا کاهش می‌یابد (سناریوی اول و دوم). با افزایش تعداد ویژگی و به دنبال آن افزایش تعداد رکوردها از ۸۱ به ۲۱۸۷، میزان خطا به طور چشمگیری کاهش می‌یابد (سناریوی اول و سوم).

نتیجه‌گیری

نتایج نشان می‌دهد با افزایش میزان آموزش، خطای درخت تصمیم کاهش و درنتیجه دقت پیش‌بینی ریسک ابتلا به سرطان سینه با استفاده از درخت تصمیم افزایش می‌یابد. در داده‌های بیولوژی به دلیل حساسیت مدل‌های پیش‌بینی‌کننده، خطای درخت تصمیم حتی با ۶۶/۶۶٪ آموزش بالا است. از طرفی با افزایش تعداد SNP درخت تصمیم از ۴ به ۷ مارکر، خطای درخت تصمیم با ۷۰/۱٪ آموزش، به‌طور چشمگیری کاهش داشت. در مجموع می‌توان گفت که با افزایش رکوردهای مجموعه آموزش و همچنین افزایش تعداد ویژگی SNP در درخت تصمیم، دقت پیش‌بینی افزایش و خطا کاهش می‌یابد. همچنین درصد شیوع بیماری در میزان خطا به دلیل انتخاب مجموعه‌های آموزش و آزمایش به روش سیستماتیک، در کد طراحی شده در این تحقیق و نرم‌افزار موجود WEKA تأثیری ندارد.

سپاسگزاری

مقاله حاضر بخشی از پایان‌نامه کارشناسی ارشد مهندسی نرم‌افزار کامپیوتر دانشگاه آزاد اسلامی واحد یزد است. بدین‌وسیله نویسندگان مقاله مراتب تشکر و قدردانی خود را اعضای محترم دانشگاه آزاد اسلامی یزد و دانشگاه جامع علمی کاربردی یزد به دلیل همکاری‌های اجرایی به عمل می‌آورند.

References:

1-Rokach L, Maimon O. ***Data mining with decision trees. theory and application***. World Scientific 2008: 1-244.

2-Ha S, Bae S, Park S. ***Web mining for distance education***. In IEEE international conference on management of innovation and technology 2000; 715–19.

3-Kingsford C, Salzberg S. ***What are decision trees?***. nature biotechnology 2008; 26(9): 1011–13.

4-Kushyar MM, Nasiri MR, Bitarafsani M, Aslaminejad AA. ***Feasibility Study of the Detection of SNPs Associated with Breast Cancer by Genome-Wide Association Virtual Studies***. J Genetic dar hezare sevom 2014; 11(3): 3190-3199. [Persian]

5-Moor J, Asselbergs F, Williams F. ***review: bioinformatics challenges for genome-wide association studies***. Genetics and population analysis 2010; 26(4): 445–55.

6-Aragones J, Ruiz J, Jimenez G, Perez J, Conejos E ***A***. ***Combined neural network and decision trees model for prognosis of breast cancer relapse***. Artificial Intelligence in Medicine 2003; 27(1): 45-63.

7-Sumbaly R, Vishnusri N, Jeyalatha S. ***Diagnosis of Breast Cancer using Decision Tree Data Mining Technique.*** International J Computer Application 2014; 98(10): 16-24.

8-Deepika M, Mary Gladence L, Madhu Keerthana R. ***A review on prediction of breast cancer using various data mining techniques***. Res J Pharmaceutical, Biological and Chemical Sci 2016; 808-14.

9-Delshi Howsalya Devi R, Indra Devi M. ***Outlier detection algorithm combined with decision tree classifier for early diagnosis of breast cancer***. International J Advanced Engineer Techno 2016; 7(2): 93-98.

10-Wang X, Peng Q, Fan Y. ***Detecting Susceptibility to Breast Cancer with SNP-SNP Interaction Using BPSOHS and Emotional Neural Networks***. Hindawi Publishing Corporation BioMed Research International 2016: 1-7.

11-Listgarten J, Damaraju S, Poulin B, Cook L, Dufour J, Driga A, et al. ***Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms.*** Clinical Cancer Res 2004; 10(8): 2725-37.

12-Ramırez N, Mesa H, Calvet H, Martınez R. ***Discovering interobserver variability in the cytodiagnosis of breast cancer using decision trees and Bayesian networks***. Applied Soft Computing 2009; 9: 1331–42.

13-Chen K, Wang K, Tsai M, Wang K, Adrian A, Cheng WC, et al. ***Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm***. BMC Bioinformatics 2014; 15(3): 1-20.

The Studies of Decision Tree in Estimation of Breast Cancer Risk by Using Polymorphism Nucleotide

Frida Seyedmir [1], Kamal Mirzaie [2], Morteza Bitaraf Sani [*3]

[1] Department of Computer Engineering, Yazd Branch, Islamic Azad University, Yazd, Iran
[2] Department of Computer Engineering, Meybod Branch, Islamic Azad University, Meybod, Iran
[3] University of Applied Sciences & Technology (Agriculture Research and Education Center), Yazd, Iran

Received: 9 Dec 2015 **Accepted:** 21 Jul 2016

Abstract

Introduction: Decision tree is the data mining tools to collect, accurate prediction and sift information from massive amounts of data that are used widely in the field of computational biology and bioinformatics. In bioinformatics can be predict on diseases, including breast cancer. The use of genomic data, including single nucleotide polymorphisms is a very important factor in predicting the risk of diseases.

Methods: By prospective analytical study **,** the risks of breast cancer were calculated associated with the use of SNP formula: $x_j = f_o * \sum_{i=1}^{m} \ln(OR_i) \times SNP_{ij}$. and decision tree. Seven SNP with different odds ratio associated with breast cancer considered; coding and designing of decision tree model, C4.5, by Csharp 2013 programming language were done.

Results: In both scenarios of coding, by increasing the training percentage from 66/66 to 86/42, the error reduced from 55/56 to 9/09. Furthermore, by running WEKA on three scenarios, including different sets of data, the number of different trainings, and different tests, the error rate decreased from 48/15 to 13/46 by increasing records number from 81 to 2187. Moreover, in the majority of scenarios, prevalence of the disease, had no effect on errors in the WEKA and code.

Conclusion: The results suggest that by increasing the amount of training decision tree error is reduced; therefore, the accuracy of the prediction of breast cancer risks through decision trees will increase. With increasing the number of records of training and increasing the number of SNP in the decision tree, the prediction accuracy will increase and errors will decrease.

Keywords: Decision Tree, Breast Cancer, Single Nucleotide Polymorphisms

This paper should be cited as:
Frida Seyedmir, Kamal Mirzaie, Morteza Bitaraf Sani. **The Studies of Decision Tree in Estimation of Breast Cancer Risk by Using Polymorphism Nucleotide.** J Shahid Sadoughi Univ Med Sci 2017; 25(4): 300-10.

***Corresponding author: Tel: 09133550060, email: bitaraf@sau.ac.ir**