

Breast Cancer Detection Using Ensemble Classifiers for Accuracy Improvement¹

Mahboubeh Shamsi²
mohadaseh karimian³
marziyeh karimian⁴

Abstract

Early diagnosis of breast cancer plays a crucial role in treating the patient. Nowadays, data mining algorithms can provide intelligent methods in the health and treatment system that accurately detect breast cancer. The purpose of this study is breast cancer detection using ensemble classifier based on WBC and WDBC prepared databases. Our proposed model in the WBC database (reducing features by cfs+ optimizing samples using Resample+ ensemble classifier using data mining algorithms (kstar + random forest + Naïve Bayes and Bayes network)) has the best detection accuracy (100%), implementation time (0 seconds) and without any errors and on the WDBC database (reducing features by cfs+ optimizing samples using Resample+ ensemble classifier using data mining algorithms (IBK algorithm+ Naïve Bayes, Bayes network and kstar)) has an accuracy of 99/29, the implementation time is 0 seconds, and the mean absolute error is 0/007. The results of this study show that according to the ensemble classifier methods using data mining algorithms on the prepared database, new systems can be designed to help physicians that facilitate treatment processes.

Keywords: Accuracy Improvement, Data Mining, Ensemble Classifiers, Feature Selection, Sampling

¹ Copyright ©the authors

² Assistant Prof. faculty of Electrical and Computer, Qom University of Technology, Qom, Irany

³ Msc. of Computer Engineering, Faculty of Electrical and Computer Engineering, Shahab Danesh University, Qom, Iran

⁴ Msc. of Computer Engineering, Faculty of Electrical and Computer Engineering, Shahab Danesh University, Qom, Iran

تشخیص سرطان سینه با استفاده از طبقه‌بندهای ترکیبی جهت بهبود دقت*

محبوبه شمسی^۱

محدثه کریمیان^۲

مرضیه کریمیان^۳

چکیده

تشخیص زودهنگام سرطان سینه نقش بسیار کلیدی در درمان بیمار ایفا می‌کند. امروزه الگوریتم‌های داده‌کاوی می‌توانند روش‌های هوشمندی در نظام سلامت ارائه دهند که با دقت بالایی سرطان سینه را تشخیص دهند. هدف از انجام این مطالعه، تشخیص سرطان سینه با استفاده از طبقه‌بندهای ترکیبی بر روی پایگاه داده‌ی آماده‌سازی شده‌ی WBC و WDBC می‌باشد. مدل پیشنهادی ما در پایگاه داده‌ی WBC (کاهش ویژگی‌ها با CFS + بهینه‌کردن نمونه‌ها با روش Resample + طبقه‌بند ترکیبی (kstar + جنگل تصادفی + شبکه‌ی بیز و بیزین ساده))، دارای بهترین دقت تشخیص (۱۰۰٪)، زمان پیاده‌سازی (۰ ثانیه) و بدون هیچ خطایی می‌باشد و در پایگاه داده‌ی WDBC (کاهش ویژگی‌ها با CFS + بهینه‌کردن نمونه‌ها با روش Resample + طبقه‌بند ترکیبی (الگوریتم IBK + شبکه‌ی بیز، بیزین ساده و kstar))، دارای دقت ۹۹/۲۹٪، زمان پیاده‌سازی ۰ ثانیه و میانگین خطای مطلق ۰/۰۰۷ می‌باشد. نتایج این مطالعه نشان می‌دهد که با توجه به روش‌های طبقه‌بند ترکیبی بر روی پایگاه داده‌ی آماده‌سازی شده می‌توان سیستم‌های نوینی برای کمک به پزشکان طراحی نمود که موجب تسهیل در فرآیندهای تشخیصی و درمانی شوند.

واژه‌های کلیدی: انتخاب ویژگی، بهبود دقت، داده‌کاوی، طبقه‌بندهای ترکیبی، نمونه‌گیری

* تاریخ دریافت: ۱۳۹۹/۰۱/۲۶؛ تاریخ پذیرش: ۱۳۹۹/۰۳/۲۲.

^۱ استادیار دانشکده برق و کامپیوتر، دانشگاه صنعتی قم، قم، ایران

^۲ کارشناسی ارشد مهندسی کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه شهاب دانش، قم، ایران

^۳ کارشناسی ارشد مهندسی کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه شهاب دانش، قم، ایران

مقدمه

سرطان سینه یکی از شایع‌ترین علت مرگ‌ومیر در زنان محسوب می‌شود و در سال‌های اخیر رشد قابل ملاحظه‌ای در تعداد افراد مبتلا به این بیماری گزارش شده است (سیگل، میلر و جمال، ۲۰۱۰). محققان میزان بالای مرگ‌ومیر زنان بر اثر سرطان سینه را ناشی از تشخیص دیر هنگام این بیماری می‌دانند. با تشخیص دقیق و به‌موقع سرطان سینه و پیشرفت‌های به‌دست آمده در درمان، می‌توان با ارائه‌ی درمان مؤثر و راهکارهای ویژه، به افزایش بقا، کاهش مرگ و ارتقا کیفیت زندگی بیماران کمک شایانی نمود (جوشی و مهتا، ۲۰۱۵). یکی از کاربردهای مهم داده‌کاوی مربوط به حوزه‌ی پزشکی و تشخیص بیماری‌ها است. امروزه به دلیل گسترش دانش در حوزه‌ی پزشکی و پیچیدگی تصمیمات مرتبط با تشخیص و درمان، توجه متخصصین به استفاده از ابزارهای هوشمند و تکنولوژی‌های کامپیوتر در امور پزشکی جلب شده است. (گبنا، کریستوفرو ویتوند، ۲۰۱۷). در این مقاله، با بررسی تکنیک‌های متفاوت یادگیری ماشین^۱ و طبقه‌بندهای ترکیبی، سیستمی کارآمد برای تشخیص سرطان سینه طراحی می‌گردد که با دقت بالایی خوش‌خیم یا بدخیم بودن تومورهای سینه را تشخیص می‌دهد. در بخش روش‌شناسی پژوهش، در ابتدا مشخصات پایگاه داده‌ها بیان می‌شوند. سپس مراحل پیش‌پردازش داده‌ها^۲ برای آماده‌سازی پایگاه داده‌ها شرح داده می‌شوند. آنگاه طبقه‌بندهای استفاده شده و فرآیندهای ساخت طبقه‌بند ترکیبی توضیح داده می‌شوند. سپس ارزیابی مدل با استفاده از مقیاس‌های کارایی بیان می‌شود. در بخش یافته‌های پژوهش، مدل پیشنهادی ارائه می‌شود و به مقایسه‌ی نتایج روش خود با نتایج روش‌های دیگران (مقالات مرتبط) پرداخته می‌شود و در بخش آخر نتیجه‌گیری و پیشنهادات بیان می‌شود. در این مقاله ابزار WEKA^۳ برای آزمایشات مورد استفاده قرار گرفته است.

پیشینه پژوهش

تاکنون تحقیقات زیادی در رابطه با تشخیص سرطان سینه با کمک الگوریتم‌های

^۱. Machine learning algorithms

^۲. Data preprocessing

^۳. WEKA (Waikato Environment for Knowledge Analysis). <http://www.cs.waikato.ac.nz/~ml/weka/>.

داده‌کاوی انجام شده است. ایلباز تأثیر طبقه‌بندهای متفاوت را در بهبود پیش‌بینی و تشخیص بیماری‌ها برای متخصصان پزشکی ارائه داد. در این مقاله برای تشخیص سریع و به‌موقع سرطان سینه، روش ترکیبی^۱ KNN+C4.5+SVM+MLP+BN با کاهش ابعاد داده^۲ Rough Set در پایگاه داده‌ی WBC پیشنهاد شد (ایلباز، ۲۰۱۵).

جیاسینگک و ولچامی با استفاده از الگوریتم‌های انتخاب ویژگی در داده‌کاوی، گامی حیاتی در افزایش دقت طبقه‌بندها در تشخیص سریع و به‌موقع سرطان سینه برداشتند. در این مقاله، روش طبقه‌بند Random Forest با استفاده از الگوریتم Modified Bat برای انتخاب ویژگی در پایگاه داده‌ی WDBC پیشنهاد شد (جیاسینگک و ولچامی، ۲۰۱۷).

جُشی و محتا با بهره‌گیری از تکنیک‌های متنوع و در دسترس یادگیری ماشین، در دو مقاله به تشخیص سرطان سینه پرداختند. در یک مقاله با مقایسه‌ی نتایج طبقه‌بندهای شبکه عصبی همچون MLP، RBF، perceptron و شبکه‌ی عصبی عمیق^۳، الگوریتم MLP با کاهش ویژگی با استفاده از الگوریتم طبقه‌بندی آنالیز گسسته‌ی خطی^۴، بهترین طبقه‌بند در پایگاه داده‌ی WDBC شناخته شد (جُشی و محتا، ۲۰۱۸ب) و در مقاله‌ی دیگر با ارائه‌ی روش پیشنهادی الگوریتم K نزدیک‌ترین همسایه^۵ با کاهش ویژگی توسط آنالیز گسسته‌ی خطی در پایگاه داده‌ی WDBC به پیش‌بینی سرطان سینه پرداختند (جُشی و محتا، ۲۰۱۸الف).

رُهان و همکاران با بررسی یک مدل طبقه‌بند ترکیبی به طبقه‌بندی انواع تومورهای خوش‌خیم و بدخیم از مجموعه داده‌های WBC پرداختند. آن‌ها با ترکیب دو الگوریتم یادگیری ماشین با ناظر از جمله جنگل تصادفی (RF) و AdaBoost بهبود عملکرد پیش‌بینی مدل را نشان دادند (روهان، سدیک، اسلام و یوسف، ۲۰۱۹).

آراچ و همکاران، با مقایسه‌ی دقت بین طبقه‌بندهای مختلف RF، BN، SMO، J48، IBK، MLP و ترکیب آن‌ها در دو پایگاه داده، نشان دادند که طبقه‌بند ترکیبی موجب بهبود دقت نتایج می‌شود. طبقه‌بند ترکیبی BN+RF+SMO+IBK به همراه انتخاب

^۱. Bayes Net

^۲. Data dimensionality reduction

^۳. Deep neural network

^۴. Linear Discriminant Analysis (LDA)

^۵. K-Nearest Neighbor (KNN)

ویژگی PCA در پایگاه داده‌ی WBC و SMO+BN، SMO+RF، SMO+MLP، SMO+IBK در پایگاه داده‌ی WDBC، بهترین نتایج را نسبت به سایر طبقه‌بندهای ترکیبی داشتند (آراچ و بادن، ۲۰۱۹).

آویناش و همکاران، عملکرد طبقه‌بند ترکیبی SVM+SMO را در پایگاه داده‌ی WBC ارزیابی کردند (آویناش، بیژوی و جی‌اراج، ۲۰۲۰).

روش‌شناسی پژوهش

در این پژوهش، ابتدا مشخصات پایگاه داده‌ها توضیح داده می‌شوند. سپس مراحل پیش‌پردازش داده‌ها برای آماده‌سازی پایگاه داده‌ها بیان می‌شوند. آنگاه طبقه‌بندهای استفاده شده و تمام فرآیندهای ساخت طبقه‌بند ترکیبی شرح داده می‌شوند. سپس ارزیابی مدل با استفاده از مقیاس‌های کارایی بیان می‌شود.

مشخصات پایگاه داده‌ها

در این مقاله از دو پایگاه داده‌ی سرطان سینه WBC و WDBC برگرفته از مخزن UCI استفاده شده است.

پایگاه داده‌ی WBC توسط دکتر ویلیام وُبرگ در سال ۱۹۹۲، ساخته شد و از دانشگاه ویسکانسین-مدیسون، ایالات متحده آمریکا ارائه شده است. پایگاه داده‌ی WBC شامل ۶۹۹ نمونه می‌باشد؛ هر نمونه شامل ۱۱ ویژگی است. ۱۰ ویژگی، اطلاعات خصیصه‌ها را نشان می‌دهند در حالی که ۱ ویژگی شامل اطلاعات کلاس است. ویژگی کلاس دو مقدار ۲=خوش‌خیم^۱ و ۴=بدخیم^۲ را نمایش می‌دهد. ۱۶ مقدار از دست‌رفته^۳ در این پایگاه داده وجود دارد که با "؟" نشان داده می‌شوند. توزیع کلاس شامل ۴۵۸ نمونه‌ی خوش‌خیم (۶۵/۵٪) و ۲۴۱ نمونه‌ی بدخیم (۳۴/۵٪) است و یک پایگاه داده‌ی ناهماهنگ^۴ است. (نیومن، هتیج، بلیک، مرز و آها، ۱۹۹۸).

^۱. Benign
^۲. Malignant
^۳. Missing values
^۴. Unbalanced

پایگاه داده‌ی WDBC به کمک دکتر ویلیام ولبرگ، نیک استریت و آلوی مانگاسارین در سال ۱۹۹۵، به وجود آمد و توسط دانشگاه ویسکانسین-مدیسون ایالات متحده آمریکا ارائه شد. پایگاه داده‌ی WDBC شامل ۵۶۹ نمونه است؛ هر نمونه دارای ۳۲ ویژگی است. ویژگی اول شامل شماره شناسه‌ی پرونده‌ی بیمار می‌باشد، ویژگی آخر، ویژگی کلاس می‌باشد و ۳۰ ویژگی باقی‌مانده از تصویر دیجیتالی آزمایش اسپیراسیون سوزنی توده‌ی پستان به دست آمده است که این ویژگی‌ها، خصوصیات هسته‌ی سلول در تصویر را بیان می‌کنند. هریک از نمونه‌های پایگاه داده‌ی WDBC با یک برچسب خوش‌خیم یا بدخیم مشخص می‌گردند. از ۵۶۹ نمونه‌ی مذکور، ۳۵۷ نمونه دارای برچسب خوش‌خیم و ۲۱۲ نمونه دارای برچسب بدخیم هستند. این پایگاه داده مقادیر از دست‌رفته ندارد و یک پایگاه داده‌ی ناهم‌بند است (نیومن، هتیچ، بلیک، مرز و آها، ۱۹۹۸).

پیش‌پردازش پایگاه داده

مرحله‌ی آماده‌سازی داده‌ها با عنوان پیش‌پردازش داده‌ها شناخته می‌شود و جهت بهبود کیفیت داده‌های واقعی برای داده‌کاوی لازم است (هان، پی و کامبر، ۲۰۱۱). برای آماده‌سازی پایگاه داده، دو مرحله‌ی پاک‌سازی داده^۱ و کاهش ابعاد داده^۲ را به ترتیب بر روی پایگاه داده‌ها انجام می‌دهیم که در ادامه این دو مرحله را شرح می‌دهیم.

پاک‌سازی داده

در این مرحله حذف یا جایگزینی مقادیر از دست‌رفته انجام می‌شود. روش استفاده‌شده در این مقاله، جایگزینی با میانگین (MS)^۳ می‌باشد. بدین صورت که داده‌های از دست‌رفته در هر ویژگی را با میانگین مقادیر همان ویژگی جایگزین می‌کند. در اکثر تحقیقاتی که مسئله‌ی داده‌های از دست‌رفته گزارش و شناسایی شده است، از روش MS برای حل این مشکل استفاده کرده‌اند (پنگ، هارول، لیو و احمان، ۲۰۰۶). بنابراین در این تحقیق نیز، از روش MS برای مقادیر از دست‌رفته استفاده می‌شود.

¹. Data cleaning
². Data dimensionality reduction
³. Mean substitution

کاهش ابعاد داده

برای کاهش ابعاد داده از سه مرحله‌ی کاهش ویژگی‌ها، بهینه کردن نمونه‌ها و تبدیل داده استفاده می‌کنیم که در ادامه این سه مرحله توضیح داده می‌شود.

الف) کاهش ویژگی‌ها (انتخاب ویژگی‌های مناسب)

به دلیل ناهمگنی^۱ پایگاه داده‌های پزشکی، افزونگی‌های زیادی در میان داده‌های با ابعاد بالا وجود دارد. حال در مقاله‌ی خود اثبات کرده است که "دقت طبقه‌بندی با استفاده از کاهش مجموعه ویژگی‌ها برابر یا بهتر از دقت طبقه‌بندی با استفاده از همه‌ی ویژگی‌ها است" (هال، ۱۹۹۹). انتخاب یک مجموعه ویژگی که خصوصیات مفید بیشتری را می‌گیرد؛ همیشه نمی‌تواند کارایی طبقه‌بندی را بهبود دهد؛ اما پیچیدگی زمانی و مکانی را کاهش می‌دهد. در این مقاله، روش انتخاب ویژگی تحت و کابری روی مجموعه داده‌ی آموزشی به کار برده می‌شود. کاهش ویژگی با استفاده از روش انتخاب ویژگی مبتنی بر همبستگی (CFS)^۲ برای مشخص کردن متریک‌های مرتبط با ویژگی کلاس استفاده شده است؛ زیرا CFS مبتنی بر متد همبستگی است و کارایی بهتری دارد. از فواید آن می‌توان، بهبود دقت تشخیص، کاهش در زمان اجرا و ابعاد بازگشتی آن را نام برد (هال، ۹۹۹؛ میکالک و کواسنیکا، ۲۰۰۶).

در پایگاه داده‌ی WBC، بعد از اعمال CFS، بهترین ویژگی‌ها که در تشخیص سرطان سینه مؤثر هستند به دست می‌آید که شامل همان ۹ ویژگی و ویژگی کلاس، می‌باشد و تنها، ویژگی شماره شناسه‌ی بیمار حذف می‌شود.

در پایگاه داده‌ی WDBC، بعد از اعمال CFS، بهترین ویژگی‌ها، شامل ۱ ویژگی Worst Concave Points و ویژگی کلاس (Diagnosis) می‌باشد و سایر ویژگی‌ها که از اهمیت کم‌تری برخوردار هستند، حذف می‌شوند.

ب) بهینه کردن نمونه‌ها

ممکن است مدلی که در پایگاه داده‌های ناهماهنگ ایجاد می‌شود، بهینه نباشد؛ زیرا

^۱. Heterogeneity

^۲. Correlation Feature Selection

بیشتر الگوریتم‌های داده کاوی، پایگاه داده را هماهنگ در نظر می‌گیرند. بنابراین، با استفاده از تکنیک‌های نمونه‌گیری، داده‌ها را قبل از استفاده از الگوریتم‌ها، هماهنگ می‌کنیم. تکنیک‌های نمونه‌گیری یا هماهنگ کردن داده‌ها به دو دسته تقسیم می‌شوند:

۱. نمونه‌گیری بیش از حد^۱: این گروه از الگوریتم‌ها با افزایش داده‌های کلاس کمینه (بدخیم)، داده‌ها را هماهنگ می‌کند.

۲. نمونه‌گیری کمتر از حد^۲: این گروه با حذف نمونه‌های کلاس اکثریت (خوش‌خیم)، داده‌ها را هماهنگ می‌کند.

روش Resample یک روش نمونه‌گیری تصادفی است که هر دو نمونه‌گیری بیش از حد و کم‌تر از حد را انجام می‌دهد (چاولا، جاپکوویچ و کوتچ، ۲۰۰۴) و مانند SMOTE یک نمونه‌گیری با ناظر^۳ است. Resample برای تولید یک نمونه‌ی تصادفی می‌تواند، داده‌های نمونه را با جایگزینی و بدون جایگزینی انجام دهد. در هنگام استفاده از این فیلتر پایگاه داده باید دارای ویژگی کلاس اسمی باشد. از آنجاکه روش Resample، هر دو نمونه‌گیری بیش از حد و کم‌تر از حد را، با جایگزینی و بدون جایگزینی، انجام می‌دهد، هم‌چنین باعث بهبود دقت طبقه‌بندها می‌شود و قابلیت اطمینان در برابر نویز را بالا می‌برد (راچمن، خودرا و ویدانتورو، ۲۰۱۷)؛ در این مقاله از این روش برای نمونه‌گیری استفاده می‌کنیم.

ج) تبدیل داده

فعالیت‌هایی مانند نرمال‌سازی، گسسته‌سازی^۴ و استانداردسازی داده‌ها، هم‌چنین تبدیل ویژگی اسمی به یک رشته^۵، دودویی^۶، تبدیل ویژگی عددی به اسمی^۷، دودویی^۸ و ... در این حوزه جای می‌گیرند.

1. Over-sampling
2. Under-sampling
3. Supervised
4. Discretization
5. Nominal To String
6. Nominal To Binary
7. Numeric To Nominal
8. Numeric To Binary

نرمال سازی مهم ترین مرحله ی تبدیل داده در نظر گرفته می شود و مجموعه ی داده ها توسط نرمال سازی تبدیل می شوند. نرمال سازی مینیمم-ماکزیمم بر فاصله ی میان نقاط حداقل و حداکثر تأکید دارد و مقیاس سازی را بر اساس تفاوت این دو مقدار انجام می دهد، که بازه ی مقادیر ویژگی ها به [۰, ۱] مقیاس می شوند. مقادیر جدید طبق رابطه ی ۱ محاسبه می شوند:

$$X^* = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \quad \text{رابطه ۱}$$

که در آن $\max(x_i)$ و $\min(x_i)$ به ترتیب مینیمم و ماکزیمم مقدار متغیر x_i است.

از آنجا که اکثر مقالات موجود در زمینه ی تشخیص سرطان سینه از نرمال سازی مینیمم-ماکزیمم (عبدالله، العزیزو الشهران، ۲۰۱۸؛ کراوسیک، ۲۰۱۵) و تبدیل ویژگی عددی به اسمی (آلیامی و همکاران، ۲۰۱۷؛ گوپتا و شالینی، ۲۰۱۸) استفاده کرده اند؛ در این مقاله نیز از روش های نرمال سازی مینیمم-ماکزیمم و تبدیل ویژگی عددی به اسمی استفاده می شود. پایگاه داده ی اصلی و پایگاه داده ی آماده سازی شده به ترتیب در جداول ۱ و ۲ نشان داده شده است.

جدول ۱. پایگاه داده ی اصلی

پایگاه داده	تعداد نمونه ها	تعداد ویژگی ها	داده های از دست رفته	تعداد کلاس ها	نوع ویژگی ها
WBC	۶۹۹	۱۱	۱۶	۲	(Integer)Numeric
WDBC	۵۶۹	۳۲	-	۲	(Real)Numeric

جدول ۲. پایگاه داده ی آماده سازی شده

پایگاه داده	تعداد نمونه ها	تعداد ویژگی ها	داده های از دست رفته	تعداد کلاس ها	نوع ویژگی ها
WBC	۶۹۹ (نرمال سازی و نمونه گیری شده)	۱۰ (نرمال سازی شده)	---	۲	Nominal
WDBC	۵۶۹ (نرمال سازی و نمونه گیری شده)	۲ (نرمال سازی شده)	---	۲	Nominal

الگوریتم‌های طبقه‌بندی

در این مقاله، با توجه به طبقه‌بندهای پر کاربرد در زمینه‌ی پزشکی و تشخیص سرطان سینه، از طبقه‌بندهای جنگل تصادفی، IBK ، K^* ، شبکه‌ی بیز، بیزین ساده^۲ و طبقه‌بند ترکیبی استفاده شده است. که در ادامه توضیح مختصری راجع به هر روش (طبقه‌بند) داده می‌شود.

جنگل تصادفی

جنگل تصادفی توسط بریمن (نتون و اولسون، ۲۰۰۰) پیشنهاد شد و جزء روش‌های یادگیری طبقه‌بند ترکیبی می‌باشد. در این روش همه‌ی طبقه‌بندها، درخت تصمیم است؛ بنابراین مجموعه‌ی همه‌ی طبقه‌بندها تشکیل جنگل را می‌دهد. هر درخت تصمیم با استفاده از انتخاب تصادفی ویژگی‌ها در هر گره انشعاب را تشخیص می‌دهد. در بین طبقه‌بندی هر درخت رأی می‌دهد و طرفدارترین رأی انتخاب می‌شود. جنگل تصادفی به تعداد ویژگی‌های انتخاب انشعاب حساس نیست؛ زیرا جنگل تصادفی از ویژگی‌های کمی در هر انشعاب استفاده می‌کند؛ پس در پایگاه‌داده‌های بزرگ مناسب است و از Bagging و Boosting (طبقه‌بندهای ترکیبی) سریع‌تر است.

الگوریتم IBK

روش K نزدیک‌ترین همسایه (KNN) یک تکنیک دسته‌بندی است که تصمیم‌گیری در مورد این که یک نمونه‌ی جدید در کدام کلاس قرار گیرد با بررسی تعدادی (k) از شبیه‌ترین نمونه‌ها یا همسایه‌ها انجام می‌شود. این روش برای یافتن شباهت بین نمونه‌ها نیاز به یک معیار فاصله نظیر فاصله‌ی اقلیدسی یا فاصله‌ی منهن دارد (هان، پی و کامبر، ۲۰۱۱). IBK یک طبقه‌بندی KNN است که از اندازه‌ی متریک استفاده می‌کند. تابع فاصله‌ی استفاده‌شده، یک پارامتر از روش جستجو است که شامل اقلیدس، چیشوف، منهن و مساحت Minkowski می‌باشد.

¹. Instance Based for K-Nearest neighbor

². Naïve Bayes (NB)

الگوریتم K* (KStar)

این الگوریتم یک طبقه‌بندی مبتنی بر نمونه است. تفاوت این طبقه‌بندی با مابقی یادگیرنده‌های مبتنی بر نمونه در استفاده از توابع فاصله است و از توابع فاصله بر اساس آنتروپی استفاده می‌کند.

روشی که برای محاسبه‌ی فاصله‌ی بین دو نمونه استفاده می‌شود از نظریه‌ی اطلاعات استفاده می‌کند. فاصله‌ی بین دو نمونه به معنای پیچیدگی انتقال یک نمونه به دیگری است. در ابتدا مجموعه‌ی متناهی از اطلاعاتی که نمونه‌ها را به نمونه‌های دیگر نگاشت می‌دهد تعریف می‌شود. یک برنامه که یک نمونه‌ی (a) را به دیگری (b) انتقال می‌دهد یک دنباله‌ی متناهی از انتقال‌هایی است که از a شروع می‌شود و به b ختم می‌شود (کلاری، تریگس، ۱۹۹۵).

بیزین ساده

این روش یکی از کارآمدترین الگوریتم‌های یادگیری می‌باشد. بر اساس فرض، استقلال قوی، یک ساختار ساده است که ارتباط بین متغیر مستقل و متغیر وابسته را آنالیز می‌کند تا یک احتمال شرطی را برای هر ارتباط استنتاج کند. با استفاده از قضیه‌ی بیز داریم:

$$P(H|X) = P(X|H) \times P(H)/P(X) \quad \text{رابطه ۲}$$

در معادله‌ی ۲، X یک رکورد داده است. H تعدادی فرض است و به رکوردی از X اشاره می‌کند که به یک کلاس خاص C تعلق دارد. $P(H|X)$ یک احتمال پسین شرطی H روی X است. $P(H)$ نیز یک احتمال پیشین است. احتمال پسین $P(H|X)$ ، مبتنی بر اطلاعات بیشتری مانند دانش پیش‌زمینه‌ای از احتمال پیشین $P(H)$ که مستقل از X است، می‌باشد. به‌طور مشابه $P(H|X)$ یک احتمال پسین شرطی X بر روی H است (ژانگ و سو، ۲۰۰۸).

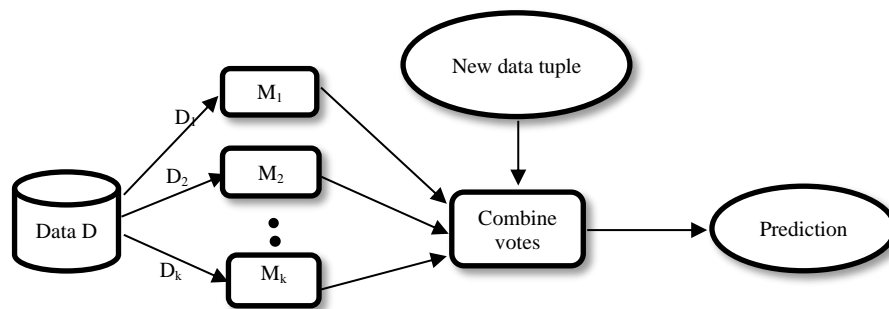
شبکه‌ی بیز

شبکه‌ی بیز (جنسنس، ۱۹۹۶) یک مدل گرافیکی احتمالی است که مجموعه‌ای از

متغیرهای تصادفی و وابستگی شرطی را از طریق گراف غیر چرخشی جهت‌دار^۱ نشان می‌دهد. به عنوان مثال، یک شبکه‌ی بیز می‌تواند رابطه‌ی احتمالی بین بیماری و علائم آن را نشان دهد. با گرفتن علائم بیماری، شبکه می‌تواند احتمال وجود انواع بیماری‌ها را بیان کند.

طبقه‌بند ترکیبی

Boosting, Bagging و جنگل تصادفی از طبقه‌بندهای ترکیبی هستند. در روش طبقه‌بند ترکیبی، یک دسته‌کتابی از مدل‌های یادگیری را باهم ترکیب می‌کند و یک مدل بهبودیافته‌ی مرکب از آن‌ها به دست می‌آورد. از یک پایگاه‌داده، k مجموعه‌ی آموزشی ایجاد می‌شود که مدل دسته‌بند M_i را می‌سازد. سپس با گرفتن یک داده‌ی جدید، هر طبقه‌بند رأی خود را می‌دهد و طبقه‌بند ترکیبی رأی‌ها را باهم ترکیب می‌کند و برچسب کلاس پیشگویی شده را برمی‌گرداند (وزنیاک، گرانا و کورچادو، ۲۰۱۴). دقت طبقه‌بند ترکیبی بیشتر از یک طبقه‌بند است؛ زیرا برفرض اگر طبقه‌بند ترکیبی اکثریت آرا^۲ را برگرداند، با دادن یک داده‌ی چندتایی، همه‌ی طبقه‌بندها برچسب کلاس پیشگویی را برمی‌گردانند و خروجی، برچسب اکثریت می‌باشد. (کیتلر، هاتف، دوین و ماتاس، ۱۹۹۸). در شکل ۱ روش طبقه‌بند ترکیبی برای ساخت مدل نشان داده شده است (وست دی، منگیامیلی، رامپال و وست وی، ۲۰۰۵).



شکل ۱. روش طبقه‌بند ترکیبی برای ساخت مدل طبقه‌بندی

^۱.Directed acyclic graph
^۲.Majority Voting

ارزیابی مدل

در این پژوهش، با توجه به مقالات بررسی شده در جهت تشخیص سرطان سینه، برای ارزیابی مدل از مقیاس‌های کارایی که مهم‌ترین آن‌ها دقت، زمان پیاده‌سازی و میانگین خطای مطلق می‌باشد، استفاده شده است. در ادامه توضیح مختصری راجع به هر کدام داده می‌شود.

دقت

دقت با نام نرخ طبقه‌بندی درست نیز شناخته شده است و نسبت تعداد سلول‌های سرطانی که درست پیش‌بینی شده‌اند به همه سلول‌ها است. این مقیاس به صورت رابطه‌ی ۳ محاسبه می‌شود (کورو و لیو، ۲۰۰۵).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (\text{رابطه ۳})$$

زمان پیاده‌سازی

هرچه یک الگوریتم در مدت‌زمان کمتری پیاده‌سازی شود، کارایی و اثربخشی بهتری دارد. زمان پیاده‌سازی به اندازه‌ی دقت طبقه‌بندیها مهم است (ایلپاز، ۲۰۱۵).

میانگین خطای مطلق

میانگین خطای مطلق (MAE)، میانگین نمونه‌ی آزمایشی است که تفاوت‌های مطلق میان مشاهدات پیش‌بینی شده و مشاهدات واقعی را نشان می‌دهد که در آن، همه‌ی تفاوت‌های منحصر به فرد، وزن یکسان دارند. در این پژوهش، از روش MAE برای اندازه‌گیری خطای استفاده شده است (چوراسیا، پال، ۲۰۱۷).

سایر معیارهای ارزیابی کارایی مدل

روش‌های متعددی جهت مقایسه‌ی طبقه‌بندی وجود دارند. کارایی پیشگویی مدل‌ها برای دو کلاس بدخیم (عود سرطان) و خوش‌خیم (عدم عود سرطان) معمولاً با استفاده از ماتریس اغتشاش که در جدول ۳ نشان داده شده است بیان می‌گردد.

جدول ۳. ماتریس اغتشاش

	NO (Prediction)	Yes (Prediction)
NO (Actual)	True Negative (TN)	False Positive (FP)
Yes (Actual)	False Negative (FN)	True Positive (TP)

ستون‌ها، نتایج پیشگویی شده و ردیف‌ها، نتایج واقعی برچسب کلاس‌ها را نشان می‌دهند. سلول‌های بدخیم با برچسب YES و سلول‌های خوش‌خیم با برچسب NO نمایش داده شده‌اند. بنابراین عناصر قطری (TN,TP) در جدول ۳ پیش‌بینی درست و بقیه (FN,FP) پیش‌بینی اشتباه را بیان می‌کنند.

درستی

درستی^۱، نسبت تعداد سلول‌های بدخیم را که درست پیش‌بینی شده است به همه سلول‌هایی که بدخیم پیش‌بینی شده است، نشان می‌دهد و به صورت رابطه ۴ محاسبه می‌شود.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (\text{رابطه ۴})$$

هرچه درستی افزایش یابد تلاش کمتری در آزمون و بررسی هدر می‌رود (کورو و لیو، ۲۰۰۵).

نرخ کشف خطا

نرخ کشف خطا^۲ مقیاسی است که نسبت تعداد سلول‌های بدخیم را که درست پیش‌بینی شده‌اند به همه سلول‌هایی که در واقع بدخیم هستند نشان می‌دهد و به صورت رابطه ۵ محاسبه می‌شود. هرچه نرخ کشف خطا افزایش یابد سلول‌های بدخیم بیشتری کشف می‌شوند (کورو و لیو، ۲۰۰۵).

$$\text{Recall} = PD = \frac{TP}{TP + FN} \quad (\text{رابطه ۵})$$

مقیاس F

مقیاس F، میانگین هارمونیک^۳ درستی و نرخ کشف خطا است که تصویر کلی خوبی از

¹. Precision
². Recall
³.harmonic

کارایی پیش‌بینی را ارائه می‌دهد (ویتن و فرانک، ۲۰۰۵) و به‌صورت رابطه‌ی ۶ محاسبه می‌شود.

$$F - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (\text{رابطه ۶})$$

ناحیه زیر منحنی ROC

مشخصات عملیاتی گیرنده^۱ (ROC) می‌تواند برای ارزیابی کارایی مدل‌های تشخیص و پیش‌بینی سرطان استفاده شود. منحنی ROC، از دو پارامتر نرخ کشف خطا (PD) در محور Y و احتمال هشدار اشتباه^۲ (PF) در محور X استفاده می‌کند و کارایی مدل را پیش‌بینی می‌کند. پارامترهای PD (Recall) و PF نیز از روی ماتریس اغتشاش محاسبه می‌شوند که به ترتیب در رابطه‌های ۵ و ۷ آورده شده‌اند.

$$PF = \frac{FP}{FP + TN} \quad (\text{رابطه ۷})$$

منحنی ROC باید از نقاط (۰,۰) و (۱,۱) عبور کند (منریز، گرین والد و فرانک، ۲۰۰۷). مهم‌ترین نواحی منحنی ROC در شکل ۲ نشان داده شده است. نقطه‌ی ایده‌آل منحنی در (۰,۱) است و هیچ خطایی در این نقطه وجود ندارد. خط گذرنده از نقاط (۰,۰) و (۱,۱) هیچ اطلاعاتی در اختیار قرار نمی‌دهد و به همین دلیل مقدار ناحیه‌ی زیر منحنی ROC (AUC^۳)، باید بیشتر از ۰/۵ باشد. اگر نتیجه روی منحنی منفی^۴ قرار بگیرد نتیجه‌ی خوبی نیست و به معنی غیرقابل قبول بودن کارایی طبقه‌بندی است. منحنی مناسب^۵ در شکل ۲ نشان داده شده است. در ناحیه‌ی مخالف هزینه^۶ احتمال کشف خطا کم‌تر است و در زمانی که بودجه‌ی اعتباریابی و واریسی کم است، مناسب است. ناحیه‌ی مخالف خطر^۷ دارای هزینه‌ی زیادی است زیرا باینکه احتمال کشف خطا زیاد است ولی احتمال هشدار خطا نیز زیاد است. برای سیستم‌های مأموریت بحرانی، ناحیه مخالف خطر و برای برنامه‌های تجارت ناحیه‌ی مخالف هزینه انتخاب می‌شود.

^۱. Receiver Operating Characteristics

^۲. Probability Of False Alarm

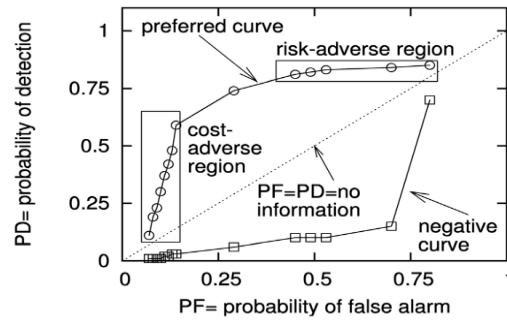
^۳. Area Under Curve

^۴. Negative curve

^۵. Preferred curve

^۶. Cost-adverse

^۷. Risk-adverse



شکل ۲. نواحی مختلف منحنی ROC (منزیز، گرین والد و فراتک، ۲۰۰۷).

یافته‌های پژوهش

در این مقاله، برای مقایسه‌ی دقیق الگوریتم‌های طبقه‌بندی، از ۲ سناریو بر روی ۲ پایگاه داده‌ی سرطان سینه (برگرفته از UCI) استفاده شده است. ۱۵ روش ساخت مدل (۵ الگوریتم طبقه‌بندی در سناریوی اول (S1) و ۱۰ الگوریتم طبقه‌بند ترکیبی در سناریوی دوم (S2)) در هر پایگاه داده بررسی شده است و ارزیابی با روش 10-fold cross validation (۱۰ بار ارزیابی) انجام شده است. اعمال روش 10-fold cross validation بر روی ۲ پایگاه داده WBC و WDBC در جدول ۴ بررسی می‌شود.

جدول ۴. اعمال روش 10-fold cross validation برای هر پایگاه داده

10-fold cross validation		تعداد نمونه‌ها	پایگاه داده
مجموعه داده آموزشی	مجموعه داده آزمایشی		
۷۰	۶۲۹	۶۹۹	WBC
۵۷	۵۱۲	۵۶۹	WDBC

نتایج نیز بر اساس مقیاس‌های کارایی از جمله دقت، زمان پیاده‌سازی و میانگین خطای مطلق بیان گردیده است. ۲ سناریو عبارتند از:

- سناریو اول (S1): انتخاب پایگاه داده‌ی آماده‌سازی شده و اعمال طبقه‌بندها به صورت جداگانه
- سناریو دوم (S2): انتخاب پایگاه داده‌ی آماده‌سازی شده و اعمال طبقه‌بندهای ترکیبی

در S1، پیاده‌سازی الگوریتم‌های طبقه‌بندی به صورت جداگانه بر روی پایگاه داده‌ها می‌باشد و در S2، طبقه‌بندی‌های ترکیبی بر روی پایگاه داده‌ها اعمال می‌شوند. در این پژوهش، روش ترکیب طبقه‌بندیها بر اساس اکثریت آرا می‌باشد و ایده‌ی ترکیب طبقه‌بندیها در هر پایگاه داده به این صورت می‌باشد که ابتدا بهترین طبقه‌بندی انتخاب می‌شود و سپس با سایر طبقه‌بندیها ترکیب می‌شود. در مرحله‌ی بعد بهترین ترکیب دوتایی طبقه‌بندیها انتخاب می‌شود، آن‌گاه با سایر طبقه‌بندیها ترکیب می‌شود. در مرحله‌ی بعد بهترین ترکیب سه‌تایی طبقه‌بندیها انتخاب می‌شود و با سایر طبقه‌بندیها ترکیب می‌شود. این کار، تا زمانی که همه‌ی طبقه‌بندیها با هم ترکیب شوند، ادامه داده می‌شود و در مرحله‌ی آخر فقط یک طبقه‌بندی وجود دارد. آن‌گاه با مقایسه‌ی همه‌ی طبقه‌بندیهای ترکیبی بهترین نتایج انتخاب می‌شود. در ادامه جداول و نمودارهای مربوط به هر پایگاه داده در هر سناریو ارائه می‌شود و توضیحات هر یک داده می‌شود.

سناریو اول: انتخاب پایگاه داده‌ی آماده‌سازی شده و اعمال طبقه‌بندیها به صورت جداگانه

در پایگاه داده‌ی WBC، ۲ طبقه‌بندی K* و RF دقت یکسانی دارند (۹۹/۸۵٪). همچنین زمان پیاده‌سازی آنها نیز یکسان (۰ ثانیه) است. اما K* بهتر است، زیرا میانگین خطای مطلق آن کمتر است (به ترتیب ۰/۰۰۳ و ۰/۰۱).

در پایگاه داده‌ی WDBC، ۲ طبقه‌بندی IBK و BN دارای دقت ۹۹/۲۹٪ و زمان پیاده‌سازی (۰ ثانیه) یکسانی دارند. اما میانگین خطای مطلق آنها متفاوت است (به ترتیب ۰/۰۰۸ و ۰/۰۰۸). بنابراین IBK بهتر از BN است، زیرا میانگین خطای مطلق آن کمتر است. مقایسه‌ی دقت طبقه‌بندیها در ۲ پایگاه داده در جدول ۵ آورده شده است.

جدول ۵. مقایسه‌ی دقت طبقه‌بندیها در ۲ پایگاه داده

BN	NB	K*	IBK	RF	طبقه‌بندیها
۹۶/۹۹	۹۷/۱۳	۹۹/۸۵	۹۹/۷۱	۹۹/۸۵	WBC
۹۹/۲۹	۹۵/۷۸	۷۹/۹۶	۹۹/۲۹	۹۹/۱۲	WDBC

تشخیص سرطان سینه با استفاده از طبقه‌بندهای ترکیبی جهت بهبود دقت // ۱۶۳

سناریو دوم: انتخاب پایگاه داده‌ی آماده‌سازی شده و اعمال طبقه‌بندهای ترکیبی
در پایگاه داده‌ی WBC، همان‌طور که در جدول ۶ نشان داده شده است، سه طبقه‌بند ترکیبی $K^*+RF+NB$ ، $K^*+RF+BN$ و $K^*+RF+NB+IBK$ بالاترین دقت را دارند (۱۰۰٪). اما، زمان پیاده‌سازی طبقه‌بند ترکیبی $K^*+RF+NB+IBK$ ، ۰/۰۲ ثانیه و دو طبقه‌بند ترکیبی دیگر ۰ ثانیه می‌باشد. بنابراین، دو طبقه‌بند ترکیبی $K^*+RF+NB$ و $K^*+RF+BN$ بهترین می‌باشند، زیرا دارای بهترین زمان پیاده‌سازی (۰ ثانیه) و بدون هیچ خطایی (میانگین خطای مطلق برابر ۰ است) می‌باشند. در پایگاه داده‌ی WBC، طبقه‌بند ترکیبی دقت بیش‌تری نسبت به طبقه‌بند تنها دارد (به ترتیب ۱۰۰٪ و ۹۹/۸۵٪). بنابراین در WBC، طبقه‌بند ترکیبی بهترین می‌باشد.

در پایگاه داده‌ی WDBC، همان‌طور که در جدول ۷ نشان داده شده است، چهار طبقه‌بند ترکیبی $IBK+BN$ ، $IBK+BN+NB$ ، $IBK+BN+K^*$ ، $IBK+BN+RF$ بالاترین دقت را دارند (۹۹/۲۹٪). در حالی که زمان پیاده‌سازی آن‌ها متفاوت است. زمان پیاده‌سازی طبقه‌بند ترکیبی $IBK+BN+RF$ ، ۰/۰۲ ثانیه و سه طبقه‌بند ترکیبی دیگر دارای بهترین زمان پیاده‌سازی می‌باشند (۰ ثانیه). همچنین میانگین خطای مطلق این سه طبقه‌بند ترکیبی باهم برابر است (۰/۰۰۷). بنابراین سه طبقه‌بند ترکیبی $IBK+BN$ ، $IBK+BN+NB$ و $IBK+BN+K^*$ بهترین می‌باشند. در سناریوی اول، طبقه‌بند ترکیبی در مقایسه با طبقه‌بند تنها با وجود یکسان بودن میزان دقت و زمان پیاده‌سازی، بهتر است زیرا میانگین خطای مطلق آن کمتر است (به ترتیب ۰/۰۰۷ و ۰/۰۰۸). بنابراین در این پایگاه داده (WDBC) نیز طبقه‌بند ترکیبی، بهترین است.

طبق نتایج به دست آمده در ۲ پایگاه داده، طبقه‌بند ترکیبی در مقایسه با طبقه‌بند تنها بهتر است و کارایی بالاتری دارد. بنابراین مدل پیشنهادی در این پژوهش، سناریوی دوم می‌باشد. در جداول ۶ و ۷ مقایسه‌ی دقت طبقه‌بندهای ترکیبی برای هر پایگاه داده ارائه شده است.

جدول ۶. مقایسه‌ی دقت در پایگاه داده WBC

K*+IBK	K*+BN	K*+NB	K*+RF	مرحله دوم
۹۹/۸۵	۹۸/۵۶	۹۸/۷۱	۹۹/۸۵	
	K*+RF+IBK	K*+RF+BN	K*+RF+NB	مرحله سوم
	۹۹/۸۵	۱۰۰	۱۰۰	
		K*+RF+NB+IBK	K*+RF+NB+BN	مرحله چهارم
		۱۰۰	۹۸/۷۱	
			K*+RF+NB+IBK+BN	مرحله پنجم
			۹۹/۸۵	

جدول ۷. مقایسه‌ی دقت در پایگاه داده WDBC

IBK+RF	IBK+K*	IBK+NB	IBK+BN	مرحله دوم
۹۹/۱۲	۹۰/۸۶	۹۷/۱۸	۹۹/۲۹	
	IBK+BN+RF	IBK+NB+N+K*	IBK+BN+NB	مرحله سوم
	۹۹/۲۹	۹۹/۲۹	۹۹/۲۹	
		IBK+BN+N+B+RF	IBK+BN+NB+K*	مرحله چهارم
		۹۹/۱۲	۹۷/۱۸	
			IBK+BN+NB+RF+K*	مرحله پنجم
			۹۹/۱۲	

مقایسه‌ی نمودارهای مقیاس‌های کارایی

در این قسمت نمودارهای مربوط به هر پایگاه داده در هر سناریو، توضیح داده می‌شود. در نمودارها، S1، بهترین طبقه‌بند در سناریوی اول (طبقه‌بند تنها) و S2 بهترین طبقه‌بند در سناریوی دوم (طبقه‌بند ترکیبی) می‌باشد.

در شکل ۳، دقت بهترین طبقه‌بند، در هر ۲ سناریو برای هر ۲ پایگاه داده نشان داده شده است. در WBC بهترین طبقه‌بند در سناریوی دوم (مدل پیشنهادی (طبقه‌بند ترکیبی))، بالاترین دقت را نسبت به بهترین طبقه‌بند در سناریوی اول (طبقه‌بند تنها) دارد و در WDBC دقت بهترین طبقه‌بند سناریوی دوم با بهترین طبقه‌بند سناریوی اول برابر است.

تشخیص سرطان سینه با استفاده از طبقه‌بندهای ترکیبی جهت بهبود دقت // ۱۶۵

در شکل ۴، میانگین خطای مطلق بهترین طبقه‌بند در هر ۲ سناریو برای هر ۲ پایگاه داده نشان داده شده است که واضح است میانگین خطای مطلق بهترین طبقه‌بند سناریوی دوم از بهترین طبقه‌بند سناریوی اول کمتر است.

در شکل‌های ۵ و ۶، سایر معیارهای ارزیابی کارایی مدل، به ترتیب، برای دو پایگاه داده‌ی WBC و WDBC نشان داده شده است که در WBC بهترین طبقه‌بند سناریوی دوم بهتر از بهترین طبقه‌بند سناریوی اول است و کارایی بالاتری نسبت به آن دارد و در WDBC بهترین طبقه‌بند سناریوی اول و دوم باهم برابر هستند.

از آنجا که زمان پیاده‌سازی بهترین طبقه‌بند برای هر ۲ سناریو در هر ۲ پایگاه داده‌ی WBC و WDBC ۰ ثانیه است از رسم کردن این نمودار صرف نظر می‌کنیم. با توجه به نتایج به دست آمده از نمودارها، سناریوی دوم بهترین سناریو است و بالاترین کارایی را دارد.

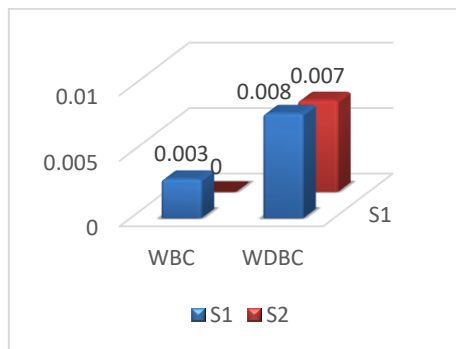
مقایسه‌ی نتایج

نتایج به دست آمده در این پژوهش، با سایر نتایج مقالات مرتبط بررسی شده است تا

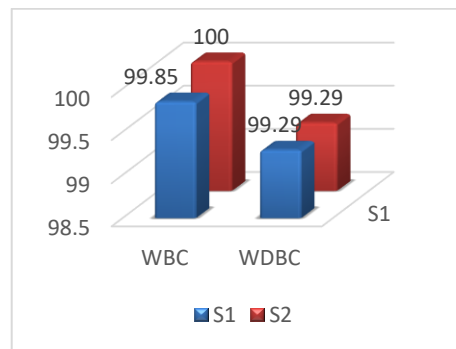
بهبود

دقت و افزایش کارایی مدل پیشنهادی نسبت به سایر مقالات مشخص شود. مقایسه‌ی

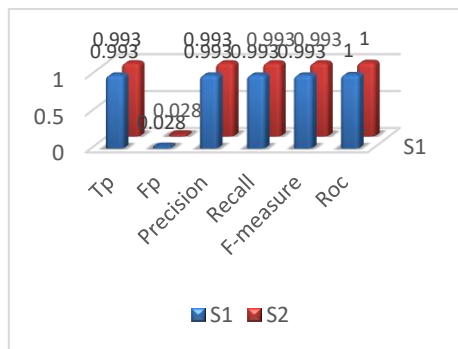
این نتایج در جدول ۸ ارائه شده است.



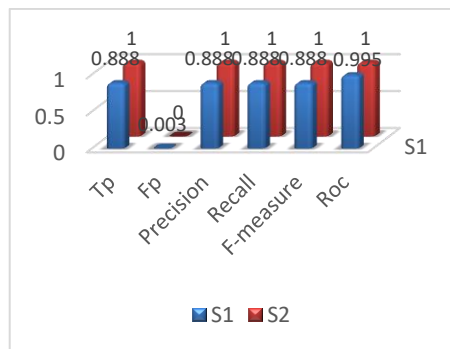
شکل ۴. ارزیابی میانگین خطای مطلق بهترین طبقه‌بند در هر سناریو در ۲ پایگاه داده



شکل ۳. ارزیابی دقت بهترین طبقه‌بند در هر سناریو در ۲ پایگاه داده



شکل ۶. ارزیابی سایر مقیاس‌های اندازه‌گیری کارایی مدل در پایگاه داده‌ی WDBC



شکل ۵. ارزیابی سایر مقیاس‌های اندازه‌گیری کارایی مدل در پایگاه داده‌ی WBC

جدول ۸. مقایسه نتایج خود با مقالات مرتبط

پایگاه داده	دقت	روش	منابع
WBC	۹۷/۲۸	SMO+J48+NB+IBK	(سلامه، عبدالحلیم و زید، ۲۰۱۲)
WBC	۹۹/۴۱	RS-KNN+C4.5+SVM+MLP+BN	(ایلباز، ۲۰۱۵)
WBC	۹۶/۸۴	SVM+RBF	(چوراسیا و پال، ۲۰۱۷ الف)
WBC	۹۹/۴۸	GA-Rotation Forest+SVM	(الیس کاویک و سویاسی، ۲۰۱۷)
WBC	۹۶/۷۱	CC-SVM+MLP	(آلیامی و همکاران، ۲۰۱۷)
WBC	۹۷	AB+RF, MBAB+RF	(آدگک، چن، بانسی و باریکزای، ۲۰۱۷)
WBC	۹۷/۱۳	SVM+NB+J48	(کومار، نیکیل و سومانگالی، ۲۰۱۷)
WBC	۹۸/۲۸	GA-SVM+RBF+Boosting	(هوانگ، چن، لین، کی و تسای، ۲۰۱۷)
WBC	۹۹/۰۴	SVM+FCM+GK	(عبدالله، العززی و الشرخان، ۲۰۱۸)
WBC	۹۸/۰۵	PCA-Rotation Forest+MLP	(آتی، خوزه، ویلسون و دیبا، ۲۰۱۸)
WBC	۹۸/۵۷	AdaBoost+RF	(روهان، سدیک، اسلام و یوسف، ۲۰۱۹)
WBC	۹۷/۹۹	PCA-BN+RF+SMO+IBK	(آراچ و بادن، ۲۰۱۹)
WBC	۹۴/۶	SVM+SMO	(آویناش، بیژوی و جیاراج، ۲۰۲۰)
WDBC	۹۷/۷۱	SMO, SMO+IBK, SMO+MLP	(سلامه، عبدالحلیم و زید، ۲۰۱۲)
WDBC	۹۵/۹۱	PCA-NB+SVM	(حازرا، ماندال و گوپتا، ۲۰۱۶)
WDBC	۹۸	PCC-Logistic Regression	(ماندال، ۲۰۱۷)
WDBC	۹۶/۸۵	Modified Bat-Random Forest	(جیاسینگ و وِجّامی، ۲۰۱۷)
WDBC	۹۲/۸	PCA-CART+EM	(نیلاشی، بن ابراهیم، احمدی و شاه‌مرادی، ۲۰۱۷)

دوفصلنامه مدیریت مهندسی و رایانش نرم
دوره ۷، شماره ۲، پاییز و زمستان ۱۴۰۰

پایگاه داده	دقت	روش	منابع
WDBC	۹۸/۵۰	Chi-Squared-MLP+Logistic	(خوریوال و میشر، ۲۰۱۸)
WDBC	۹۷/۰۶	LDA-MLP	(جُشی و محتا، ۲۰۱۸ ب)
WDBC	۹۷/۰۶	LDA-KNN	(جُشی و محتا، ۲۰۱۸ الف)
WDBC	۹۷/۷۱	SMO+RF,SMO+BN,SMO+IBK,SMO+MLP	(آراچ و بادن، ۲۰۱۹)
WBC	۱۰۰	Resample-CFS-K*+RF+BN Resample-CFS-K*+RF+NB	روش پیشنهادی
WDBC	۹۹/۲۹	Resample-CFS-IBK+BN Resample-CFS-IBK+BN+NB Resample-CFS-IBK+BN+K*	روش پیشنهادی

نتیجه‌گیری و پیشنهادها

در این مطالعه، ما با استفاده از روش طبقه‌بند ترکیبی (تجمعی)، کاهش ویژگی (با استفاده از روش انتخاب ویژگی مبتنی بر همبستگی (CFS))، نمونه‌گیری (با استفاده از روش نمونه‌گیری با ناظر Resample) و نرمال‌سازی داده‌ها، دقت شناسایی سیستم‌های سرطان سینه را افزایش دادیم. در واقع با استفاده از الگوریتم‌های داده‌کاوی می‌توان سیستم‌های نوین و با صرفه‌تری در نظام سلامت و درمان ارائه داد که با دقت بالایی قادر به تشخیص سرطان سینه باشند. استفاده از این سیستم‌ها می‌تواند موجب کاهش خطاهای احتمالی شود و دقت تشخیص سرطان سینه را بهبود بخشد. مدل پیشنهادی در پایگاه داده‌ی WBC، دارای بهترین دقت تشخیص (۱۰۰٪)، زمان پیاده‌سازی (۰ ثانیه) و بدون هیچ خطایی می‌باشد و در پایگاه داده‌ی WDBC، دارای دقت ۹۹/۲۹٪ زمان پیاده‌سازی ۰ ثانیه و میانگین خطای مطلق ۰/۰۰۷ می‌باشد، که در هر دو پایگاه داده نسبت به سایر روش‌ها دارای بهترین نتایج است. همچنین نتایج مطالعات محققان دیگر بر روی پایگاه داده‌های WBC و WDBC در مقایسه با روش پیشنهادی، حاکی از برتری روش پیشنهادی این مطالعه است. نتایج این مطالعه نشان می‌دهد که با توجه به روش‌های طبقه‌بند ترکیبی با استفاده از الگوریتم‌های داده‌کاوی بر روی پایگاه داده‌ی آماده‌سازی شده می‌توان سیستم‌های نوینی برای کمک به پزشکان طراحی نمود که موجب تسهیل در فرآیندهای تشخیصی و درمانی شوند.

منابع

- Abdullah, M., Al-Anzi, F., & Al-Sharhan, S. (2018). Hybrid Multistage Fuzzy Clustering System for Medical Data Classification. *Computing Sciences and Engineering (ICCSE), 2018 International Conference On*, 1-6. IEEE.
- Adegoke, V. F., Chen, D., Banissi, E., & Barikzai, S. (2017). Prediction of breast cancer survivability using ensemble algorithms. *Smart Systems and Technologies (SST), 2017 International Conference On*, 223-231. IEEE.
- Alickovic, E., & Subasi, A. (2017). Breast cancer diagnosis using GA feature selection and Rotation Forest. *Neural Computing and Applications*, 28(4), 753-763.
- Alyami, R., Alhajjaj, J., Alnajrani, B., Elaalami, I., Alqahtani, A., Aldhafferi, N., ... Olatunji, S. O. (2017). Investigating the effect of Correlation based Feature Selection on breast cancer diagnosis using Artificial Neural Network and Support Vector Machines. *Informatics, Health & Technology (ICIHT), International Conference On*, 1-7. IEEE.
- Ani, R., Jose, J., Wilson, M., & Deepa, O. S. (2018). Modified Rotation Forest Ensemble Classifier for Medical Diagnosis in Decision Support Systems. In *Progress in Advanced Computing and Intelligent Engineering* (pp. 137-146). Springer.
- Arach, S., & Bouden, H. (2019). Performance Analysis on Three Breast Cancer Datasets using Ensemble Classifiers Techniques. *Computer Science*, 14(4), 935-952.
- Avinash, K., Bijoy, M. B., & Jayaraj, P. B. (2020). Early Detection of Breast Cancer Using Support Vector Machine With Sequential Minimal Optimization. In *Advanced Computing and Intelligent Engineering* (pp. 13-24). Springer.
- Chaurasia, V., & Pal, S. (2014). Data mining techniques: to predict and resolve breast cancer survivability. *International Journal of Computer Science and Mobile Computing IJCSMC*, 3(1), 10-22.
- Chaurasia, V., & Pal, S. (2017a). *Data mining techniques: To predict and resolve breast cancer survivability*.
- Chaurasia, V., & Pal, S. (2017b). *Performance analysis of data mining algorithms for diagnosis and prediction of heart and breast cancer disease*.
- Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1), 1-6.
- Cleary, J. G., & Trigg, L. E. (1995). K*: An Instance-based Learner Using an Entropic Distance Measure. *ICML*, 108-114.
- El-Baz, A. H. (2015). Hybrid intelligent system-based rough set and ensemble classifier for breast cancer diagnosis. *Neural Computing and Applications*, 26(2), 437-446.
- Fenton, N. E., & Ohlsson, N. (2000). Quantitative analysis of faults and failures in a complex software system. *Software Engineering, IEEE Transactions On*, 26(8), 797-814.
- Gbenga, D. E., Christopher, N., & Yetunde, D. C. (2017). Performance Comparison of Machine Learning Techniques for Breast Cancer Detection. *Nova*, 6(1), 1-8.
- Gupta, P., & Shalini, L. (2018). Analysis of Machine Learning Techniques for Breast Cancer Prediction. *International Journal Of Engineering And Computer Science*, 7(05), 23891-23895.
- Hall, M. A. (1999). *Correlation-based feature selection for machine learning*.
- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
- Hazra, A., Mandal, S. K., & Gupta, A. (2016). Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms. *International Journal of Computer Applications*, 145(2).
- Huang, M.-W., Chen, C.-W., Lin, W.-C., Ke, S.-W., & Tsai, C.-F. (2017). SVM and SVM ensembles in breast cancer prediction. *PloS One*, 12(1), e0161501.
- Jensen, F. V. (1996). *An introduction to Bayesian networks* (Vol. 210). UCL press London.
- Joshi, A., & Mehta, A. (2018a). ANALYSIS OF K-NEAREST NEIGHBOR TECHNIQUE FOR BREAST CANCER DISEASE CLASSIFICATION. *Machine Learning*, 98, 13.
- Joshi, A., & Mehta, A. (2018b). BREAST CANCER DATA CLASSIFICATION USING NEURAL NETWORK AND DEEP NEURAL NETWORK TECHNIQUES. *Int J Recent Sci Res*, 9(4), 25788-25792.
- Khuriwal, N., & Mishra, N. (2018). Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm. *2018 IEEMA Engineer Infinite Conference (ETechNxt)*, 1-5. IEEE.
- Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions*

- on *Pattern Analysis and Machine Intelligence*, 20(3), 226–239.
- Koru, A. G., & Liu, H. (2005). Building effective defect-prediction models in practice. *Software, IEEE*, 22(6), 23–29.
 - Krawczyk, B. (2015). One-class classifier ensemble pruning and weighting with firefly algorithm. *Neurocomputing*, 150, 490–500.
 - Kumar, U. K., Nikhil, M. B. S., & Sumangali, K. (2017). Prediction of breast cancer using voting classifier technique. *Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2017 IEEE International Conference On*, 108–114. IEEE.
 - Mandal, S. K. (2017). Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree. *International Journal Of Engineering And Computer Science*, 6(2).
 - Menzies, T., Greenwald, J., & Frank, A. (2007). Data mining static code attributes to learn defect predictors. *Software Engineering, IEEE Transactions On*, 33(1), 2–13.
 - Michalak, K., & Kwasnicka, H. (2006). Correlation-based feature selection strategy in neural classification. *Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference On*, 1, 741–746. IEEE.
 - Newman, D. J., Hettich, S., Blake, C. L., Merz, C. J., & Aha, D. W. (1998). UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine, CA. 1998 of Conference, [Http://Archive. Ics. Uci. Edu/MI/Datasets. Html](http://Archive.Ics.Uci.Edu/MI/Datasets.Html).
 - Nilashi, M., bin Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). An analytical method for diseases prediction using machine learning techniques. *Computers & Chemical Engineering*, 106, 212–223.
 - Peng, C.-Y. J., Harwell, M., Liou, S.-M., & Ehman, L. H. (2006). Advances in missing data methods and implications for educational research. *Real Data Analysis*, 3178.
 - Rachman, G. H., Khodra, M. L., & Widyantoro, D. H. (2017). Rhetorical Sentence Categorization for Scientific Paper Using Word2Vec Semantic Representation. *Journal of Physics: Conference Series*, 801(1), 12070. IOP Publishing.
 - Rohan, T. I., Siddik, A. B., Islam, M., & Yusuf, M. S. U. (2019). A Precise Breast Cancer Detection Approach Using Ensemble of Random Forest with AdaBoost. *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, 1–4. IEEE.
 - Salama, G. I., Abdelhalim, M., & Zeid, M. A. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. *Breast Cancer (WDBC)*, 32(569), 2.
 - Siegel, R. L., Miller, K. D., & Jemal, A. (2017). Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, 67(1), 7–30.
 - Teh, Y.-C., Tan, G.-H., Taib, N. A., Rahmat, K., Westerhout, C. J., Fadzli, F., ... Yip, C.-H. (2015). Opportunistic mammography screening provides effective detection rates in a limited resource healthcare system. *BMC Cancer*, 15(1), 405.
 - West, D., Mangiameli, P., Rampal, R., & West, V. (2005). Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application. *European Journal of Operational Research*, 162(2), 532–551.
 - Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
 - Wozniak, M., Grana, M., & Corchado, E. (2014). A survey of multiple classifier systems as hybrid systems. *Information Fusion*, 16, 3–17.
 - Zhang, H., & Su, J. (2008). Naive Bayes for optimal ranking. *Journal of Experimental & Theoretical Artificial Intelligence*, 20(2), 79–93.

استناد به این مقاله:

شناسه دیجیتال (DOI): 10.22091/jemsc.2020.5425.1129

شمسی، محبوبه؛ کریمیان، محدثه، کریمیان، مرضیه. (۱۳۹۸). «تشخیص سرطان سینه با استفاده از طبقه‌بندهای ترکیبی جهت بهبود دقت». *مدیریت مهندسی و رایانش نرم*، ۷(۲)، ۱۶۹–۱۴۵.