مقاله پژوهشی

فصلنامهٔ بیماری‌های پستان ایران، سال نهم، شماره دوم، تابستان ۱۳۹۵

طبقه‌بندی داده‌های نامتوازن در تشخیص اولیه بیماری‌های پستان با روش‌های آدابوست، شبکه عصبی احتمالی و K تا نزدیک‌ترین همسایه

محمد درزی*: گروه پژوهشی سیستم‌های اطلاعاتی پیشرفته، پژوهشکده فناوری اطلاعات و ارتباطات جهاد دانشگاهی، تهران، ایران
آسیه الفت بخش: گروه پژوهشی بیماری‌های پستان، مرکز تحقیقات سرطان پستان جهاد دانشگاهی، تهران، ایران
سعید گرگین: گروه فناوری اطلاعات و سامانه‌های هوشمند، سازمان پژوهش‌های علمی و صنعتی ایران، تهران، ایران
فرید اویسی: گروه پژوهشی سیستم‌های اطلاعاتی پیشرفته، پژوهشکده فناوری اطلاعات و ارتباطات جهاد دانشگاهی، تهران، ایران
عصمت‌السادات هاشمی: گروه پژوهشی بیماری‌های پستان، مرکز تحقیقات سرطان پستان جهاد دانشگاهی، تهران، ایران
نسرین‌السادات علوی: گروه پژوهشی بیماری‌های پستان، مرکز تحقیقات سرطان پستان جهاد دانشگاهی، تهران، ایران

چکیده

مقدمه: سرطان پستان یکی از سرطان‌های شایع در ایران بوده و هرگونه اقدام تشخیصی به هنگام در این مورد می‌تواند جان بسیاری از مبتلایان به این سرطان را نجات بخشد. هدف از این پژوهش طبقه‌بندی داده‌های نامتوازن مربوط به بانوان مراجعه‌کننده به کلینیک پژوهشکده سرطان پستان جهاددانشگاهی به منظور تعیین وضعیت ایشان و طبقه‌بندی نرمال و یا غیرنرمال بودن پستان مراجعه‌کنندگان بود. مجموعه داده‌های نامتوازن یکی از چالش‌های پیش روی طراحی سیستم‌های پزشک‌یار برای طبقه‌بندی و تعیین وضعیت بیمار محسوب می‌شود که در این پژوهش از روش‌های سطح داده برای حل آن استفاده شد.

روش بررسی: در این مطالعه برای طبقه‌بندی داده‌های ۹۱۸ نفر، سه الگوریتم AdaBoost.M1، k تا نزدیک‌ترین همسایه و شبکه عصبی احتمالی به خدمت گرفته شد. از آنجا که داده‌های این مطالعه نامتوازن بود، برای حل این مساله از روش بیش نمونه‌برداری تصادفی کلاس اقلیت، زیرنمونه برداری تصادفی کلاس اکثریت و بیش نمونه‌برداری مصنوعی کلاس اقلیت استفاده شد. به منظور پیاده‌سازی الگوریتم‌ها از امکانات و ابزارهای نرم‌افزار «متلب» و «آر» استفاده گردید. همچنین برای ورودی الگوریتم‌های طبقه‌بندی از ۶۰ متغیر مندرج در کاربرگ‌های شرح حال و معاینه فیزیکی مراجعان استفاده شد. معیارهای دقت و F-measure به منظور ارزیابی در مرحله آزمون الگوریتم‌ها مورد استفاده قرار گرفت.

یافته‌ها: بر اساس معیارهای دقت و F-measure، بهترین عملکرد الگوریتم‌های سه‌گانه این مطالعه در مواجهه با مجموعه داده تولیدشده با روش بیش نمونه‌برداری مصنوعی کلاس اقلیت بود. در این راستا عملکرد الگوریتم‌های AdaBoost.M1، k تا نزدیک‌ترین همسایه و شبکه عصبی احتمالی در مواجهه با مجموعه داده مذکور و بر اساس معیارهای دقت و F-measure به ترتیب عبارتند از: ۹۳/۵ و ۹۳/۶، ۷۹/۵ و ۸۷/۷ و ۸۶ و ۹۱/۹ بدست آمد.

نتیجه‌گیری: روش‌های مختلفی برای حل مساله عدم توازن مجموعه داده‌ها به منظور طبقه‌بندی وجود دارد؛ نمونه‌گیری مجدد که از روش‌های سطح داده محسوب می‌شود یکی از متداول‌ترین آنهاست. از سه روش نمونه‌گیری مجددی که در این مطالعه استفاده شد، بهترین عملکرد طبقه‌بندها در مواجهه با مجموعه داده ایجاد شده در نتیجه نمونه‌گیری مجدد به روش بیش نمونه‌برداری مصنوعی کلاس اقلیت بود. از بین الگوریتم‌های به خدمت گرفته شده و بر اساس معیارهای دقت و F-measure بهترین عملکرد در تمامی مجموعه داده‌های این مطالعه متعلق به الگوریتم AdaBoost.M1 بود.

واژه‌های کلیدی: عدم توازن داده، طبقه‌بندی، بیماری پستان، AdaBoost.M1، k تا نزدیک‌ترین همسایه، شبکه عصبی احتمالی، نمونه‌گیری مجدد.

* نشانی نویسنده پاسخگو: تهران، چهارراه کالج، کوچه شهید سعیدی، پلاک ۵، پژوهشکده فناوری اطلاعات و ارتباطات جهاددانشگاهی، محمد درزی.
نشانی الکترونیک: modarzi@yahoo.com

[DOR: 20.1001.1.17359406.1395.9.2.1.5]

مقدمه

از دیدگاه بسیاری از افراد، بیماری پستان مساوی با سرطان پستان است. در حالی‌که بیماری‌ها و ضایعات مرتبط با پستان موارد زیادی را شامل می‌شوند که تشخیص آنها توسط پزشکان عمومی، متخصصان زنان و جراحان لازم و ضروری است. شایع‌ترین این بیماری‌ها عبارتند از تغییرات فیبروکیستیک، کیست‌ها، فیبروآدنوم و سایر توده‌های خوش‌خیم، ترشحات نوک پستان، ماستیت، پاژه نوک پستان و بالاخره سرطان پستان.

شایع‌ترین علت مراجعه زنان به کلینیک‌های پستان، درد و سپس علایمی مثل احساس توده، ترشح از نوک پستان، احساس سفتی یا ندولاریته، تغییر رنگ یا اندازه پستان است؛ این علایم اکثراً به علت بیماری‌های خوش‌خیم پستان ایجاد می‌شوند. پزشک با جمع‌بندی اطلاعات مربوط به شرح حال و معاینه بیمار، تشخیص احتمالی را مطرح و بر این اساس اقدامات پاراکلینیک مناسب را درخواست می‌کند که ممکن است شامل ماموگرافی، سونوگرافی، MRI، آسپیراسیون سوزنی، بیوپسی سوزنی، سیتولوژی ترشحات و غیره باشد (۱).

توده پستان یک علامت مهم بیماری‌های پستان، اعم از بیماری‌های خوش‌خیم و بدخیم پستان است. شایع‌ترین علامت سرطان پستان لمس توده یا توده‌های متعدد در پستان است که معمولا بدون درد بوده و به بافت اطراف و گاهی به پوست چسبندگی دارد. دیگر علایم سرطان پستان عبارتند از ترشح از نوک پستان، تغییرات پوستی شامل قرمزی، تورفتگی یا ادم پوست پستان، زخم یا توکشیدگی نوک پستان، بزرگی غدد لنفاوی زیر بغل و با شیوع کمتر تورم بازو و اندام فوقانی و علایم درگیری ارگان‌های دیگر (۲).

کشف یک توده در پستان می‌تواند از مهمترین اتفاقات زندگی یک زن باشد که باعث اضطراب می‌شود. برخلاف تومورهای خوش‌خیم پستان، توده‌های بدخیم در صورت عدم تشخیص و درمان به موقع، بی‌وقفه رشد می‌کنند، به بافت‌های اطراف و حتی به نقاط دوردست گسترش می‌یابند و می‌توانند در صورت عدم درمان منجر به مرگ شوند.

مقابله با بیماری‌ها دارای دو بخش پیش‌گیری و درمان است. در مورد سرطان، از آنجا که در اغلب موارد علت مشخصی برای بیماری یافت نمی‌شود، پیش‌گیری مفهوم دیگری می‌یابد. امروزه دانشمندان علوم بهداشتی به واژه جدیدی به نام غربالگری دست یافته‌اند که هدف آن تشخیص زودرس، کاهش مرگ و میر، کاهش ناتوانی‌های ناشی از بیماری و عوارض درمان است. اساس غربالگری در سرطان پستان، تشخیص این بیماری در زمانی است که توده کوچک بوده و هنوز به قسمت‌های دیگر بدن انتشار پیدا نکرده است. در مراحل اولیه، ممکن است بیمار کاملاً بدون علامت باشد و پس از مراجعه به منظور کنترل، شک به وجود سرطان ایجاد شود و یا به‌وسیله روش‌های تشخیصی مثل ماموگرافی و یا سونوگرافی این شک تایید گردد. بدیهی است در این موارد ضایعه سرطانی در مراحل اولیه تشخیص داده شده و درمان آن با موفقیت بیشتری همراه خواهد بود (۱).

مهم‌ترین روش‌های تشخیص زودرس یا غربالگری در سرطان پستان عبارتند از:

۱- معاینه ماهانه پستان توسط خود فرد (خودآزمایی پستان)

۲- معاینه پستان توسط پزشک

۳- ماموگرافی

بنابراین سیستمی به عنوان پزشک‌یار برای پیش بینی و تشخیص زودرس بیماری پستان می‌تواند فرایند غربالگری را تسهیل نماید به طوری‌که در صورت غیرنرمال بودن وضعیت مراجعه‌کنندگان، اقدامات تشخیصی دقیق‌تری روی ایشان انجام شود. بر همین اساس پژوهشی با همکاری دو گروه پژوهشی سیستم‌های اطلاعاتی پیشرفته پژوهشکده فناوری اطلاعات جهاد دانشگاهی و گروه پژوهشی بیماری‌های پستان پژوهشکده سرطان پستان جهاد دانشگاهی طراحی شد و داده‌های مربوط به شرح حال و معاینه فیزیکی تعدادی از مراجعه‌کنندگان به کلینیک پژوهشکده سرطان پستان جهاد دانشگاهی به منظور طراحی سیستم مذکور توسط متخصصان مجرب و از طریق سیستمی مبتنی بر شبکه جمع‌آوری شد. یکی از ویژگی‌های عمومی داده‌های پزشکی، نامتوازن بودن آن است (۳).

با توجه به تقسیم شرایط پستان مراجعه‌کنندگان به نرمال و غیرنرمال (وضعیت خوش‌خیم و مشکوک به بدخیم)، مجموعه داده ایجاد شده در این پژوهش نیز از این مساله مستثنی نبود. مجموعه داده نامتوازن بر اساس تعریف عبارت است از یک مجموعه داده‌ای که تعداد نمونه‌های

[DOR: 20.1001.1.17359406.1395.9.2.1.5]

متعلق به یک کلاس در آن با تعداد نمونه‌های کلاس دیگر به طور مساوی توزیع نشده باشد (۴).

کلاس با تعداد داده‌های بیشتر را کلاس اکثریت و کلاس با داده‌های کمتر را کلاس اقلیت می‌نامند. در الگوریتم‌های طبقه‌بند استاندارد، توزیع کلاس‌ها متوازن در نظر گرفته می‌شود و این دسته از الگوریتم‌ها در مواجهه با مجموعه داده‌های نامتوازن عملکرد مناسبی را از خود ارایه نمی‌دهند؛ چرا که الگوریتم‌های معمول طبقه‌بند به سمت نمونه‌های آموزشی کلاس بزرگ‌تر متمایل می‌شوند که این موضوع باعث افزایش خطا در شناسایی نمونه‌های اقلیت می‌شود (۵). این مساله یکی از چالش‌های پیش رو برای طبقه‌بندی[1] داده‌های نامتوازن محسوب می‌شود و امروزه نظر بسیاری از متخصصان و پژوهشگران حوزه تحلیل داده را به خود جلب کرده است (۶). از روش‌های متنوعی برای حل مساله عدم توازن در علم یادگیری ماشین استفاده می‌شود (۷). یکی از این روش‌ها، روش‌های بازبینی در سطح الگوریتم[2] است که با تغییر در الگوریتم طبقه‌بندی به نوعی مساله عدم توازن مرتفع می‌شود (۸). از دیگر روش‌های حل نامتوازن بودن داده‌ها روش های مبتنی بر ترکیب طبقه‌بندها است. هدف اصلی روش ترکیب[3] تلاش برای بهبود عملکرد طبقه‌بندی داده‌ها از طریق ترکیب چندین طبقه‌بند است. به طوری‌که ترکیب چند طبقه‌بند عملکرد بهتری نسبت به یکی از همان طبقه‌بندها خواهد داشت. یان و همکاران با استفاده از این روش و ماشین بردار پشتیبان[4] توانستند مساله پیش‌بینی کلاس اقلیت را بهبود دهند (۹). روش سوم روش‌های سطح داده[5] است. در این دسته از روش‌ها توزیع کلاس نامتوازن با نمونه‌گیری مجدد[6] در فضای داده‌ها متوازن می‌شود (۱۰). و نهایتاً روش‌های حساس به هزینه[7] دسته دیگری از روش‌های ارایه شده برای حل عدم توازن در داده‌ها محسوب می‌شود. این دسته از روش‌ها به نوعی از ترکیب روش‌های تغییر در الگوریتم طبقه‌بند و روش‌های سطح داده حاصل می‌شوند (۱۱). در این بین، روش‌های سطح داده با دو رویکرد زیرنمونه‌برداری[8] و بیش‌نمونه‌برداری[9] از روش‌های موثر در متوازن نمودن داده‌ها محسوب می‌شوند (۱۲).

در این مطالعه با توجه به نامتوازن بودن داده‌ها از روش‌های سطح داده شامل زیرنمونه‌برداری تصادفی کلاس اکثریت[10]، بیش‌نمونه‌برداری تصادفی کلاس اقلیت[11] و بیش‌نمونه‌برداری مصنوعی کلاس اقلیت[12] برای افزایش دقت طبقه‌بندی الگوریتم‌های AdaBoost.M1، k تا نزدیک‌ترین همسایه و شبکه عصبی احتمالی در تشخیص کلاس‌های (مراجعه‌کنندگان) نرمال از غیرنرمال استفاده شد.

مواد و روش‌ها

مجموعه داده:

داده‌های این مجموعه مربوط به مراجعان کلینیک پژوهشکده سرطان پستان جهاد دانشگاهی بود که کاربرگ‌های شرح حال و معاینه فیزیکی برای ۹۱۸ نفر از مراجعه‌کنندگان به این مرکز، توسط سه تن از جراحان خبره و عضو هیات علمی جهاد دانشگاهی پس از ویزیت ایشان، ورود داده شد. این مجموعه داده پس از استخراج از پایگاه داده با نظر خبرگی ایشان مورد بررسی قرار گرفت و با نظر ایشان تعداد محدودی از رکوردها که دارای مقدار گم‌شده[13] بود، حذف شد. سپس داده‌ها برای استفاده، نرمال‌سازی شدند. عناوین ویژگی‌های مربوط به شرح حال[14] و معاینه فیزیکی[15]مراجعه‌کنندگان در جدول شماره ۱ درج شده است.

[1] Classification
[2] Algorithm level
[3] Ensemble Methodology
[4] Support Vector Machine
[5] Data Level
[6] Re-sampling
[7] Cost-sensitive learning

[8] Under-Sampling
[9] Over-Sampling
[10] random majority under sampling(RUS)
[11] random minority oversampling(ROS)
[12] Synthetic Minority Oversampling Technique (SMOTE)
[13] Missing Value
[14] History
[15] PhysicalExam

DOR: 20.1001.1.17359406.1395.9.2.1.5]

جدول ۱: ویژگی‌های مربوط به شرح حال و معاینه فیزیکی مراجعه‌کنندگان

	گروه	ردیف	عنوان متغیر	مقادیر قابل انتخاب	نوع ویژگی
ویژگی‌های مربوط به شرح حال	Reproductive status	1.	Birthday (Age)	date of birth	continuous
		2.	Menarche age	year of Menarche	continuous
		3.	Menstrual Cycle	Regular Irregular	categorical
		4.	Reproductive Status	Premenopause Perimenopause Postmenopausal	categorical
		5.	Marital Status	Single Married Widow/Divorce	categorical
		6.	Gravity	No child Gravid Infertile	categorical
		7.	abortion	Yes No	Categorical(binary)
		8.	Lactation	Month of lactating	continuous
	History of hormone Therapy	9.	HRT	Yes No	Categorical(binary)
		10.	Infertility treatment	Yes No	Categorical(binary)
		11.	OCP use	Yes No	Categorical(binary)
	Smoking	12.	Patient’s smoking	Yes No	Categorical(binary)
		13.	Family member’ Smoking	Yes No	Categorical(binary)
	Family History of Cancer	14.	Male Breast Cancer	Yes No	Categorical(binary)
		15.	Ovarian cancer	Yes No	Categorical(binary)
		16.	Uterus cancer	Yes No	Categorical(binary)
		17.	colon cancer	Yes No	Categorical(binary)
	Personal history of cancer	18.	Breast cancer	Yes No	Categorical(binary)
		19.	Ovarian cancer	Yes No	Categorical(binary)
	Other breast diseases	20.	Breast trauma	Yes No	Categorical(binary)
		21.	Breast infection	Yes No	Categorical(binary)
	Paraclinic	22.	Mammography	birads 0 birads 1 birads 2/3 birads 4 birads 5	categorical

[DOR: 20.1001.1.17359406.1395.9.2.1.5]

		23.	Sonography	birads 0 birads 1 birads 2/3 birads 4 birads 5	categorical
ویژگی‌های مربوط به معاینه فیزیکی	Chief Complaint	24.	ChiefComplaint	pain mass screening discharge skin change axillary imaging abnormality	categorical
		25.	Screening	Yes No	Categorical(binary)
		26.	Imaging abnormality	Yes No	Categorical(binary)
		27.	Pain	Yes No	Categorical(binary)
		28.	Mass	Yes No	Categorical(binary)
		29.	Duration of Mass	Month	continuous
		30.	Discharge	Yes No	Categorical(binary)
		31.	Skin Symptom	Yes No	Categorical(binary)
		32.	Nipple-Areolla complex changes	Yes No	Categorical(binary)
		33.	Asymmetry	Yes No	Categorical(binary)
	Previous Treatment	34.	Previous related Treatment	Yes No	Categorical(binary)
		35.	Response To Treatment	Yes No	Categorical(binary)
	Physical exam (PE)-Inspection	36.	Normal	L and R Normal =1 L or R Normal=2 L and R Abnormal=3	Categorical
		37.	Larger	R&L Large=0 R|L Large = 1	Categorical(binary)
		38.	Erythema	(R&L=0) = 0 (R | L=1) = 1	Categorical(binary)
		39.	Edema	(R&L=0) = 0 (R | L=1) = 1	Categorical(binary)
		40.	Peaud Orange	(R&L=0) = 0 (R | L=1) = 1	Categorical(binary)
		41.	Ulcer	(R&L=0) = 0 (R | L=1) = 1	Categorical(binary)
		42.	Dimpling	(R&L=0) = 0 (R | L=1) = 1	Categorical(binary)
		43.	Nipple Eczema	(R&L=0) = 0 (R | L=1) = 1	Categorical(binary)
		44.	Nipple Retraction	(R&L=0) = 0 (R | L=1) = 1	Categorical(binary)
		45.	Nipple Inversion	(R&L=0) = 0 (R | L=1) = 1	Categorical(binary)
		46.	Bulging	(R&L=0) = 0 (R | L=1) = 1	Categorical(binary)

DOR: 20.1001.1.17359406.1395.9.2.1.5]

	Physical exam-Palpation	47.	Normal	(R&L=0) = 0 (R | L=1) = 1	Categorical(binary)
		48.	Thickenning	(R&L=0) = 0 (R | L=1) = 1	Categorical(binary)
		49.	Nodularity	(R&L=0) = 0 (R | L=1) = 1	Categorical(binary)
		50.	Tenderness	(R&L=0) = 0 (R | L=1) = 1	Categorical(binary)
		51.	Mass	Single Multiple	Categorical(binary)
		52.	Movable	Yes No	Categorical(binary)
		53.	Fixed skin	Yes No	Categorical(binary)
		54.	Fixed deep	Yes No	Categorical(binary)
		55.	Mass Size	<2 cm 2-5 cm >5 cm	Categorical
		56.	Mass Consistency	Hard Firm Soft	Categorical
	PE -Lymph Node	57.	Axilla LN	(R&L=0) = 0 (R | L=1) = 1	Categorical(binary)
		58.	Supraclavicular LN	(R&L=0) = 0 (R | L=1) = 1	Categorical(binary)
		59.	Adhesion LN	(R&L=0) = 0 (R | L=1) = 1	Categorical(binary)
		60.	Axillary Tenderness	(R&L=0) = 0 (R | L=1) = 1	Categorical(binary)

نمونه‌گیری مجدد[16]:

مجموعه داده اولیه نامتوازن بود به طوری‌که داده‌های متعلق به کلاس نرمال- کلاس اقلیت[17]- ۶۷ رکورد و داده‌های متعلق به کلاس غیرنرمال- کلاس اکثریت[18]- ۸۵۱ رکورد بود که بر این اساس مجموعه داده این مطالعه دارای نرخ عدم توازن[19] ۱۲/۷٪ بود.

در این راستا، در گام نخست سه روش طبقه‌بندی بر روی مجموعه داده نامتوازن اصلی اجرا شد. در گام دوم مجموعه داده شماره یک با روش زیرنمونه برداری تصادفی کلاس اکثریت[20] ایجاد شد. بدین صورت که از کلاس اکثریت یک نمونه تصادفی به اندازه داده‌های کلاس اقلیت انتخاب شد تا هر دو کلاس به یک اندازه مشاهده شوند (۱۳). در گام بعد، طبقه‌بندی داده‌ها بر روی مجموعه داده شماره دو که با روش بیش‌نمونه‌برداری تصادفی کلاس اقلیت [21] ایجاد شد، اجرا گردید. دراین روش از کلاس اقلیت به صورت تصادفی نمونه‌هایی انتخاب و به همین مجموعه اضافه شد تا اندازه داده‌های کلاس اقلیت با اندازه داده‌های کلاس اکثریت برابر شود.

همچنین مجموعه داده سوم با روش بیش‌نمونه‌برداری مصنوعی کلاس اقلیت[22] بر روی مجموعه داده اصلی ایجاد شد (۴).

اطلاعات مجموعه داده اصلی و ۳ مجموعه داده جدید که بر اساس نمونه‌گیری مجدد از مجموعه داده اصلی ایجاد شد، در جدول ۲ آمده است.

[16] Re-sampling
[17] Minority
[18] Majority
[19] Imbalanced Rate
[20] random majority under sampling(RUS)
[21] random minority oversampling(ROS)
[22] Synthetic Minority Oversampling Technique (SMOTE)

[DOR: 20.1001.1.17359406.1395.9.2.1.5]

جدول ۲: مجموعه داده اصلی و سه مجموعه داده نمونه‌گیری مجدد شده از آن

مجموعه داده	سایز داده	داده نرمال	داده غیرنرمال	نسبت داده غیرنرمال به نرمال
اصلی نامتوازن	۹۱۸	۶۷	۸۵۱	۸۵۱/۶۷
شماره یک	۸۰	۴۰	۴۰	۱
شماره دو	۱۵۸۸	۷۳۷	۸۵۱	۸۵۱/۷۳۷
شماره سه	۹۵۱	۴۶۹	۴۸۲	۴۸۲/۴۶۹

روش‌های طبقه‌بندی:

AdaBoost.M1 نخستین توسعه از الگوریتم Adaboost محسوب می‌شود (۱۴). این الگوریتم که توسط یاو فروند و رابرت شاپیر ابداع شد، یکی از ۱۰ الگوریتم برتر داده کاوی محسوب می‌شود (۱۵).

شبکه عصبی احتمالی[۲۳]:

این شبکه‌ها بر اساس استراتژی بیزین و تخمین‌زننده‌های غیرپارامتریک توابع چگالی احتمال طبقه‌بندی می‌کنند. با دسترسی به داده‌های مشخص، ابزارهای قدرتمندی برای شناخت و طبقه‌بندی الگوها با بیشترین احتمال موفقیت محسوب می‌شوند (۱۶). از مزایای این الگوریتم می‌توان به حساس نبودن به داده‌های دورافتاده[۲۴] و دقت بالاتر نسبت به شبکه‌های عصبی دیگر مانند پرسپترون[۲۵] را اشاره نمود.

K تا نزدیک‌ترین همسایه[۲۶]:

این الگوریتم از دسته الگوریتم‌های یادگیری بر پایه مثال[۲۷] بوده و جزء ۱۰ الگوریتم برتر داده‌کاوی محسوب می‌شود (۱۵). در روش‌های یادگیری بر پایه مثال، فقط مثال‌ها ذخیره می‌شوند و هرگونه تعمیم تا مشاهده مثال جدید به تعویق می‌افتد. در این روش فرض می‌شود که تمام مثال‌ها نقاطی در فضای n بعدی حقیقی هستند و همسایه‌ها برمبنای فواصل اقلیدسی استاندارد تعیین می‌گردند. منظور از k تعداد همسایه‌های در نظر گرفته شده برای تعیین همسایگی مثال جدید است.

ارزیابی نتایج:

صحت[۲۸] به عنوان یک شاخص ارزیابی برای عملکرد طبقه‌بندها استفاده می‌شود.

ماتریس درهم ریختگی[۲۹]:

این ماتریس از ابزارهای مناسب برای بررسی میزان موفقیت و کارایی سیستم‌های طبقه‌بندی محسوب می‌شود. در جدول شماره ۳ پیکره اصلی آن ارایه شده است (۱۷).

در این ماتریس ۴ سنجه به شرح ذیل خواهیم داشت:

- TP_{rate}[۳۰]: میزان نمونه‌های مثبت درست. درصدی از نمونه‌های مثبت که درست طبقه‌بندی شده‌اند.
- FPrate[۳۱]: میزان نمونه‌های مثبت کاذب. درصدی از نمونه‌های مثبت که نادرست طبقه‌بندی شده‌اند.
- TNrate[۳۲]: میزان نمونه‌های منفی درست. درصدی از نمونه‌های منفی که درست طبقه‌بندی شده‌اند.
- FNrate[۳۳]: میزان نمونه‌های منفی کاذب. درصدی از نمونه‌های منفی که نادرست طبقه‌بندی شده‌اند.

جدول شماره ۳: ماتریس در هم ریختگی

		کلاس‌های پیش‌بینی شده	
		کلاس= نرمال	کلاس= غیرنرمال
کلاس‌های واقعی	کلاس= نرمال	TP	FN
	کلاس= غیرنرمال	FP	TN

[23] Probabilistic neural network
[24] Outlier Data
[25] perceptron
[26] K-Nearest Neighbor (K-NN)
[27] Instance Based Learning
[28] Accuracy
[29] Confusion matrix
[30] True Positive
[31] Fulse Positive
[32] True Negative
[33] Fulse Negative

DOR: 20.1001.1.17359406.1395.9.2.1.5]

از ایجاد این ماتریس می‌توان معیارهای ذیل را محاسبه نمود:

$$\text{صحت}: \frac{TP+TN}{TP+FP+TN+FN} \quad (۱۸)$$

$$\text{دقت}: \frac{TP}{TP+FN} \quad (۱۹)$$

$$\text{حساسیت}: \frac{TP}{TP+FN} \quad (۲۰)$$

$$F-Measure=\frac{2.\text{Re}call.\text{Pr}ecision}{\text{Re}call+precision} \quad (۲۱)$$

میانگین هندسی:

معیار بعدی برای اندازه‌گیری کارایی طبقه‌ها، میانگین هندسی(Geometric mean) است؛ ایده اصلی این معیار حداکثرسازی دقت بر روی هر دو کلاس است (۲۲). این معیار به شکل زیر تعریف می‌شود:

$$G-Mean=\sqrt{\text{Pr}ecision.\text{Re}call}$$

ابزار شبیه سازی: کلیه شبیه‌سازی‌ها در نرم افزار مَتلب انجام شد. تنها برای اجرای الگوریتم SMOTE از نرم افزار آر(R) استفاده شد. در پکیج DMwR نرم افزار آر، این الگوریتم پیاده‌سازی شده است که هم‌زمان با ساخت داده‌های مصنوعی برای ایجاد توازن در داده‌های اقلیت مجموعه داده، امکان زیر نمونه‌برداری از داده‌های کلاس اکثریت را برای کاربر فراهم می‌نماید.

یافته‌ها

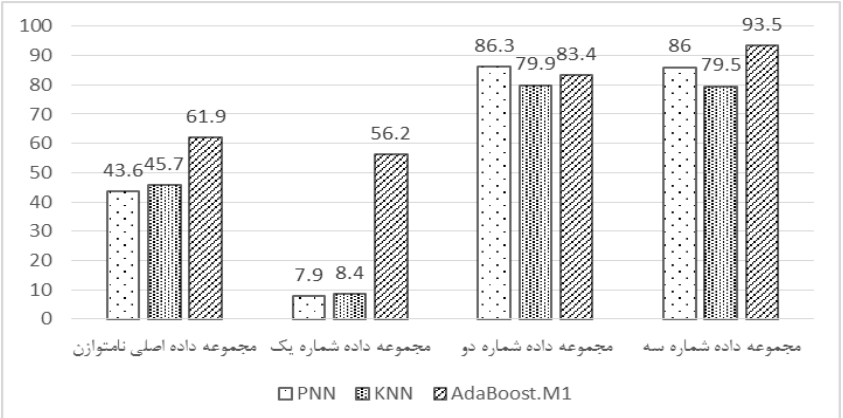
برای تولید داده‌های آموزشی و آزمون، ۱۰ بار به طور تصادفی، تعدادی از داده‌ها برای آموزش انتخاب و باقی‌مانده داده‌ها برای آزمون انتخاب شد. نمودارهای ۱، ۲، ۳ و ۴ به ترتیب دقت، صحت، میانگین هندسی و F-measure و عملکرد طبقه‌بندها در ۱۰ مرتبه تکرار است.

بر اساس نمودار ۱، بیشترین دقت عملکرد برای الگوریتم‌های سه‌گانه در مواجهه با مجموعه داده شماره سه است که با روش بیش نمونه‌برداری مصنوعی کلاس اقلیت تولید شد. در بین طبقه‌بندهای موجود، بهترین دقت متعلق به AdaBoost.M1 است.

همان‌گونه که در نمودار ۴ آمده، در بین روش‌های نمونه‌گیری مجدد به منظور رفع مساله عدم توازن مجموعه داده، بالاترین مقدار F-Measure به رفتار طبقه‌بندها با روش بیش نمونه‌برداری مصنوعی کلاس اقلیت تعلق داشته و کمترین مقدار F-Measure متعلق به زیرنمونه‌برداری کلاس اکثریت است که دلیل اصلی آن می‌تواند تعداد پایین نمونه‌ها (۴۰ رکورد) باشد. مشخصا تعداد نمونه‌های متعلق به هر کلاس از تعداد ویژگی‌ها (۶۰ ویژگی) پایین‌تر است.

به منظور استنباط بهتر نتایج این مطالعه و برای ارزیابی عملکرد الگوریتم‌های AdaBoost.M1، k تا نزدیکترین همسایه و شبکه عصبی احتمالی، میانگین یکایک معیارهای صحت، دقت، F-measure و میانگین هندسی به ازای چهار مجموعه داده مورد مطالعه، محاسبه شد و نتیجه آن در نمودار ۵ ارایه شد. براین اساس، الگوریتم AdaBoost.M1 در تمامی معیارها، عملکرد بهتری نسبت به دو روش دیگر داشت؛که این عملکرد مربوط به نوع طراحی این طبقه‌بند است.

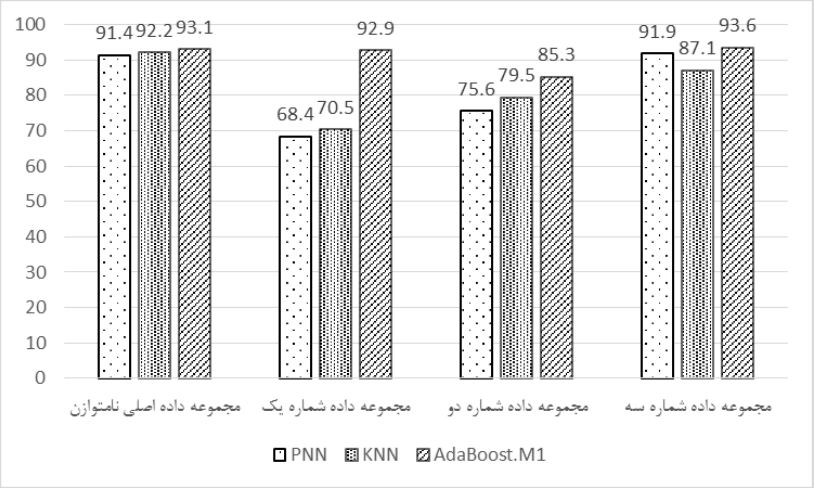
همچنین به منظور ارزیابی روش‌های طبقه‌بندی به خدمت گرفته شده در این مقاله، از تحلیل واریانس استفاده شد. بر این اساس، مقدار پی (P-Value) کمتر از 10^{-10} بدست آمد.



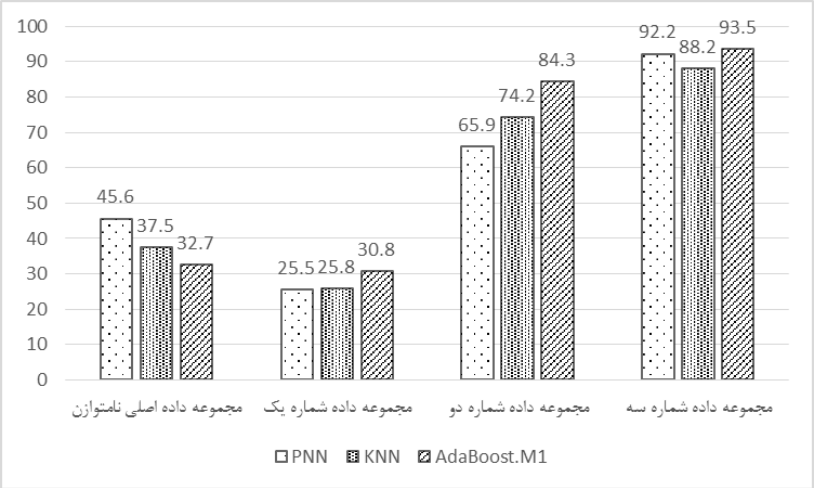
نمودار ۱: مقایسه معیار دقت الگوریتم‌های سه‌گانه طبقه‌بند در مواجهه با چهار مجموعه داده مورد مطالعه

[DOR: 20.1001.1.17359406.1395.9.2.1.5]

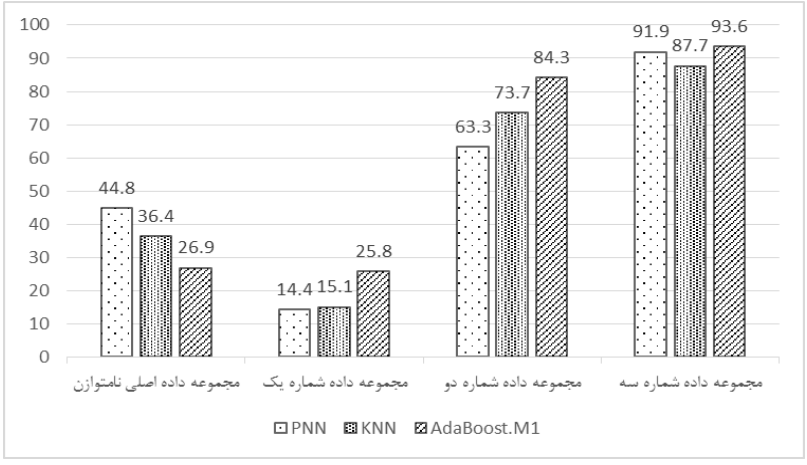
طبقه‌بندی داده‌های نامتوازن در تشخیص اولیه بیماری‌های پستان با روش‌های آدابوست، شبکه



نمودار ۱: مقایسه معیارصحت عملکرد الگوریتم‌های سه‌گانه طبقه‌بند در مواجهه با چهار مجموعه داده مورد مطالعه

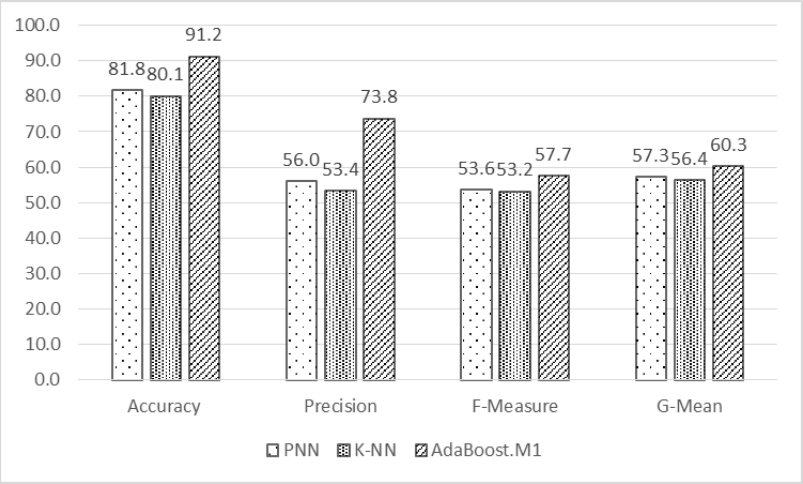


نمودار ۲: مقایسه معیار میانگین هندسی عملکرد الگوریتم‌های سه‌گانه طبقه‌بند در مواجهه با چهار مجموعه داده مورد مطالعه



نمودار ۳: مقایسه معیار F-measure عملکرد الگوریتم‌های سه‌گانه طبقه‌بند در مواجهه با چهار مجموعه داده مورد مطالعه

DOR: 20.1001.1.17359406.1395.9.2.1.5]



نمودار ۵: مقایسه مقدار میانگین معیارهای ارزیابی عملکرد الگوریتم‌ها در ۴ مجموعه داده

بحث

در این مطالعه با استفاده از سه الگوریتم AdaBoost.M1، شبکه عصبی احتمالی و k تا نزدیکترین همسایه داده‌های مربوط به مراجعه‌کنندگان کلینیک سرطان پستان که دارای مساله عدم توازن بود، طبقه‌بندی شد. با توجه به مجموعه داده‌های ایجاد شده به‌وسیله سه روش نمونه‌گیری مجدد از بین الگوریتم‌های به خدمت گرفته شده و بر اساس معیارهای دقت و F-measure بهترین عملکرد در تمامی مجموعه داده‌های این مطالعه در تعیین کلاس مراجعه‌کنندگان با وضعیت نرمال و غیرنرمال متعلق به الگوریتم AdaBoost.M1 بود.

امروزه کاوش در مجموعه داده‌های نامتوازن یکی از مسایل مهم تحلیل داده محسوب می‌شود. و عدم توازن در بین داده‌های متعلق به کلاس اقلیت و اکثریت باعث تمایل الگوریتم‌های طبقه‌بندی به سمت کلاس اکثریت می‌شود.

روش‌های نمونه‌گیری مجدد از مجموعه داده نامتوازن اصلی یکی از متداول‌ترین روش‌ها برای حل مساله عدم توازن به شمار می‌آید.

در این مطالعه سه روش نمونه‌گیری مجدد شامل زیرنمونه‌برداری تصادفی کلاس اکثریت، بیش‌نمونه‌برداری تصادفی کلاس اقلیت و بیش نمونه‌برداری مصنوعی کلاس اقلیت استفاده شدکه باعث بهبود عملکرد الگوریتم‌های مورد مطالعه در این پژوهش شد.

دو روش بیش نمونه‌برداری تصادفی کلاس اقلیت و بیش‌نمونه‌برداری مصنوعی کلاس اقلیت که در این مطالعه مورد استفاده قرار گرفت، هر دو متعلق به رویکرد بیش-نمونه‌برداری بودند. با مقایسه نتایج حاصل از ارزیابی الگوریتم‌های طبقه‌بندی این مطالعه می‌توان دریافت که این الگوریتم‌ها در مواجهه با مجموعه داده ایجاد شده به روش بیش‌نمونه‌برداری مصنوعی کلاس اقلیت عملکرد بهتری را ارایه می‌نمایند. علت در این است که اصولا در روش بیش‌نمونه‌برداری تصادفی کلاس اقلیت، تعدادی نمونه به طور تصادفی انتخاب می‌شوند و به دفعات تکرار می‌شوند؛ بر این اساس ناحیه تصمیم مربوط به داده‌های اقلیت که در طبقه‌بندی آنها توسط طبقه‌بند موثر است به طور خاص تنها محدود به فضای نمونه‌هایی می‌شود که به طور تصادفی انتخاب و تکرار شده‌اند. اما مجموعه داده ایجاد شده با روش بیش‌نمونه‌برداری مصنوعی کلاس اقلیت به طبقه‌بند اجازه می‌دهد که در مرحله تصمیم‌گیری فضای بیشتری از داده‌های اقلیت را مدنظر قرار دهد. به عبارت بهتر روش بیش‌نمونه‌برداری مصنوعی کلاس اقلیت نمونه‌های مرتبط با کلاس اقلیت را بهتر و با سطح پوشش بیشتری تولید می‌نماید که این مساله باعث آموزش[1] بهتر طبقه‌بندها و در نهایت عملکرد بهتر الگوریتم طبقه‌بندی می‌شود.

نکته دیگر مورد بحث در این مطالعه مربوط به معیارهای ارزیابی است؛ از این منظرباید گفت معیار صحت در پیش‌بینی عملکرد طبقه‌بندها در مواجهه با داده‌های نامتوازن، معیار مناسبی محسوب نمی شود. بر اساس نمودار ۲ به خوبی می‌توان ناکارآمدی این معیار در

[1] Learning

[DOR: 20.1001.1.17359406.1395.9.2.1.5]

مجموعه داده‌های نامتوازن را مشاهده نمود؛ به طوری‌که اگر چه مقدار این معیار برای هر سه طبقه‌بند بالای ۹۰٪ است ولی مقدار دقت و F-Measure در آنها پایین است. در این راستا، معیارهای دیگری را برای این دسته از مسایل که اصولا در آنها با دو نوع خطا مواجه هستیم، باید در نظر گرفت:

۱- تشخیص نادرست فرد مبتلا به بیماری پستان به عنوان فرد نرمال

۲- تشخیص اشتباه فرد نرمال به عنوان فرد بیمار

از بین این دو، اشتباه اول هزینه بالاتری برای سیستم داشته و باعث عدم درمان فرد مبتلا به بیماری می‌شود. بر این اساس معیار مناسب برای تشخیص میزان خطای سیستم در تشخیص نادرست فرد مبتلا به سرطان به عنوان فرد نرمال، استفاده از معیار دقت و F-measure است.

در مجموع استفاده از سه طبقه‌بند با سه رویکرد متفاوت الگوریتمی و استفاده از سه تن از متخصصان بیماری‌های پستان در جمع‌آوری داده‌های مراجعه‌کنندگان، از نقاط قوت این مطالعه محسوب می‌شود. همچنین به روز بودن مشخصات کاربرگ‌های شرح حال و معاینه فیزیکی مراجعه‌کنندگان که ویژگی‌های مورد بررسی در این مطالعه از آنها استخراج شد از دیگر نقاط قوت این تحقیق است. در راستای حل مساله عدم توازن، استفاده از سه روش‌نمونه‌گیری مجدد در ایجاد مجموعه داده‌های جدید نیز از نقاط قابل توجه این پژوهش کاربردی محسوب می‌شود.

در این راستا حذف تعداد محدودی از رکوردها به خاطر وجود مقادیر گم شده و عدم جایگزینی آنها یکی از محدودیت‌های این مطالعه محسوب می شود؛ چرا که هر چه حجم داده‌های مورد بررسی در مسایل طبقه‌بندی بیشتر باشد، نتایج پژوهش نیز دقیق‌تر خواهد بود. همچنین جمع‌آوری داده‌های مراجعه‌کنندگان صرفاً به یک مرکز به نوعی محدودیت محسوب می‌شود. چنانچه اگر این امکان فراهم می‌شد که داده‌های مراجعه‌کنندگان به چند کلینیک بیماری‌های پستان در یک مجموعه داده تجمیع شود، می‌توانست جامعیت مجموعه داده را افزایش دهد.

نتیجه‌گیری

روش‌های مختلفی برای حل مساله عدم توازن مجموعه داده‌ها به منظور طبقه‌بندی وجود دارد؛ نمونه‌گیری مجدد که از روش‌های سطح داده محسوب می شود یکی از متداول‌ترین آنهاست. از سه روش نمونه‌گیری مجددی که در این مطالعه استفاده شد، بهترین عملکرد طبقه‌بندها در مواجهه با مجموعه داده ایجاد شده در نتیجه نمونه‌گیری مجدد به روش بیش نمونه‌برداری مصنوعی کلاس اقلیت بود. از بین الگوریتم‌های به خدمت گرفته شده و بر اساس معیارهای دقت و F-measure بهترین عملکرد در تمامی مجموعه داده‌های این مطالعه در تعیین کلاس مراجعه‌کنندگان با وضعیت نرمال و غیرنرمال متعلق به الگوریتم AdaBoost.M1 بود.

پیشنهادات

به کارگیری تعداد بیشتری از الگوریتم‌های طبقه‌بندی، به کارگیری روش‌های دیگر حل مساله عدم توازن، و استفاده از روش‌های انتخاب ویژگی از جمله پیشنهادات نویسندگان برای انجام پژوهش‌های آتی در این زمینه است.

تقدیر و تشکر

در اینجا لازم است از زحمات آقایان دکتر حبیب ا... اصغری رئیس محترم پژوهشکده ICT جهاد دانشگاهی، مهندس علی اصغرلیائی، مهندس سالار محتاج، دکتر محمود طاووسی عضو هیات عملی پژوهشکده علوم بهداشتی جهاد دانشگاهی، سرکار خانم دکتر معصومه مداح عضو محترم هیات علمی پژوهشکده ICT جهاد دانشگاهی، و دکتر مهدی سادات رسول عضو هیات علمی دانشگاه خوارزمی به خاطر حمایت و هم‌فکری ایشان در تدوین مقاله تقدیر و تشکر نماییم.

References

۱. حقیقت شهپر و دیگران. آشنایی با بیماری‌های پستان. تهران: سازمان انتشارات جهاد دانشگاهی،۱۳۹۳.

2. Harris JR, Lippman ME, Morrow M, Osborne CK. Diseases of the breast. 3rd

DOR: 20.1001.1.17359406.1395.9.2.1.5]

ed. Philadelphia: Lippincott Williams and Wilkins 2004.

3. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. J Intelli Data Analysis 2002; 5:429-49.
4. Chawla NV. Data Mining for Imbalanced Datasets: An Overview. Data mining know discov handbook 2005.
5. Sun Y, Wong AKC, Kamel MS. Classification of Imbalancd Data: A Review. Int J Patt Recogn Artif Intell 2009; 4:687-719.
6. Yang Q, WU X.10 Challenging Problems in Data Mining Research. IntJInf Tech& Dec Mak 2006; 4:597-604.
7. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A Review on Ensembles for the Class Imbalance Problem: Bagging- Boosting- and Hybrid-Based Approaches. IEEE Trans on Syst Man Cyber Part C AppRevi 2012; 4:463-84.
8. Barandela R, S´anchez JS, Garc´ıa V, Rangel E. Strategies for learning in class imbalance problems. Patt Recogn 2003; 3:849-51.
9. Yan R, Liu Y, Jin R, Hauptmann A. On predicting rare classes with SVM ensembles in scene classification. Proceeding of IEEE International Conference on on Acoustics, Speech and Signal Processing 2003; 6(10): III-21–III-24.
10. Napierała K, Stefanowski j, Wilk S. Learning from Imbalanced data in presence of noisy and borderline examples. In: Szczuka M, Kryszkiewicz M, Ramanna S, Jensen R, Hu Q, editors. RSCTC, LNAI 6086. Proceedingof 7th International Conference; 2010June 28-30; Warsaw, Poland. 2010. p.158–167.
11. Zhang S, Liu L, Zhu X, Zhang C. A strategy for attributes selection in cost-sensitive decision trees induction. Proceeding of IEEE 8th International Conference onComputer and Information Technology Workshops 2008; 8(11): 8-13.
12. Li DC, Liu CW, Hu SC. A learning method for the class imbalance problem with medical data sets. J Comput Bio Medi 2010; 5: 509-18.
13. Rahman MM, Davis DN. Addressing the Class Imbalance Problem in Medical Datasets. Int J Machine Learning and Comput 2013; 2: 224-8.
14. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 1997; 1:119-39.
15. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. J Knowl Inf Syst 2007; 1:1–37.
16. Wasserman P. Advanced Methods in Neural Computing. New York: Van Nostrand Reinhold 1993.
17. Powers DMW. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness and Markedness & Correlation. J Machine Learning Tech 2011; 1: 37-63.
18. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-Sampling Technique. J. Artifi Intelli Research 2002; 16: 321-57.
19. Guo H, Viktor HL. Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost IM Approach. ACM SIGKDD Explorations Newsletter 2004; 1: 30-9.
20. Woods K, Doss C, Bowyer K, Solka J, Priebe C, Kegelmeyer W. Comparative Evaluation of Pattern Recognition Techniques for Detection of Microcalcifications in Mammography. Int J Patt Recogn Artifi Intelli 1993; 6: 1417-36.
21. Rao RB, Krishnan S, Niculescu RS. Data Mining for Improved Cardiac Care. ACM SIGKDD Explorations Newsletter 2006; 1: 3-10.
22. Estabrooks A. A Combination Scheme for Inductive Learning from Imbalanced Data Sets.MCS Thesis, Dalhousie University 2000.

[DOR: 20.1001.1.17359406.1395.9.2.1.5]