فصلنامه علمی تخصصی ایده های نو در علوم، مهندسی و فناوری ، دوره ۷، شماره ۱، بهار ۱۴۰۲، ص ۵۷-۵۰
ISSN: 2588-3984
http://www.Tajournals.com

عنوان مقاله: دل‌سازی اولیه بیماری پارکینسون در داده های نا متوازن با استفاده از داده کـاوی و تشخیص اولیه بیماری
دوره ۷/ شماره ۱ / بهار ۱۴۰۲/ ص ۵۷- ۵۰

مدل‌سازی اولیه بیماری پارکینسون در داده های نا متوازن با استفاده از داده کـاوی و تشخیص اولیه بیماری

سعیده یار محمدلو
کارشناسی ارشد مهندسی صنایع – دانشگاه الزهرا- تهران - ایران

چکیده

در این مقاله به بررسی مدل‌سازی اولیه بیماری پارکینسون در داده های نا متوازن با استفاده از داده کـاوی پرداخته شد. بیماری پارکینسون اختلالی است که متاثر از سلول‌های عصبی در قسمتی از مغز که در ارتباط با حرکات است می باشد. افراد مبتلا به این بیماری اغلب تجربههایی همچون لرزش، سفتی ماهیچه‌ها، اشکال در راه رفتن، مشکلات تعادلی و حرکات آهسته دارند. تشخیص این بیماری مشکل و پرهزینه است و تشخیص زود هنگام این بیماری نیاز به یک روش تشخیص صحیح و مطمئن است تا آنرا از سایر بیماری‌های مشابه تمیز دهند. بنابراین، یافتن یک روش تشخیص صحیح و موثر و همچنین عوامل خطر در بروز این بیماری، بسیار با اهمیت است. بطور کلی ترکیب طبقه‌بندها یک زمینه تحقیقاتی جدید در مبحث یادگیری ماشین و تشخیص الگو می‌باشد. مدل‌سازی روش‌های ترکیبی به این صورت است که مجموعه‌ای از طبقه‌بندها را با داده-های آموزشی ایجاد کرده و میزان صحت را با انجام عملیات رای‌گیری بر روی نتایج آن‌ها بدست می‌آورند. از طرفی با وجود داده‌های نامتوازن و یادگیری در مجموعه داده‌های نامتوازن جایی که نمونه‌های طبقه اکثریت خیلی بیشتر از بقیه است، چالش مهمی در یادگیری ماشین است زیرا الگورتیم‌های قدیمی یادگیری ماشین، ممکن است به سمت طبقه اکثریت متمایل شوند و این مسئله صحت پیش‌بینی را در طبقه اقلیت پایین می‌آورد. در این مقاله روش مدل‌سازی زیرنمونه برداری تصادفی را بعد از مقایسه آن با روش‌های دیگر نمونه‌برداری مانند بیش‌نمونه‌برداری تصادفی وEasyEnsemble و Bagging Modified ، برای نمونه‌برداری مجموعه آموزش استفاده شد و سپس نتایج را با معیارهای Recall و Precision و معیار F و معیار G ارزیابی نموده تا توان پیش-بینی طبقه‌بندها در مقابل داده‌های نامتوازن افزایش پیدا کند. در این مقاله روشی را بر مبنای استخراج ویژگی بر مبنای طبقه‌بند ترکیبی شد تا داده‌های آموزش را برای طبقه‌بند پایه ایجاد شود.

کلمات کلیدی: مدل‌سازی طبقه بند ترکیبی، ماشین بردار پشتیبان، بیماری پارکینسون، طبقه بندی.

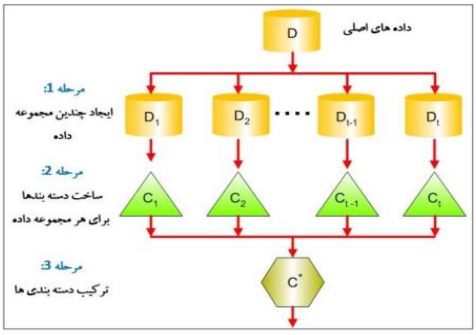
فصلنامه علمی تخصصی ایده های نو در علوم، مهندسی و فناوری ، دوره ۷، شماره ۱، بهار ۱۴۰۲، ص ۵۷-۵۰
ISSN: 2588-3984
http://www.Tajournals.com

مقدمه

بیماری پارکینسون اختلالی است که متاثر از سلول‌های عصبی در قسمتی از مغز که در ارتباط با حرکات است می باشد. افراد مبتلا به این بیماری اغلب تجربههایی همچون لرزش، سفتی ماهیچه‌ها، اشکال در راه رفتن، مشکلات تعادلی و حرکات آهسته دارند. این نشانه‌ها معمولا بعد از سن ۶۰ سالگی رخ می‌دهد، هر چند که برخی افراد با این بیماری جوان‌تر از ۵۰ سال هستند. بیماری پارکینسون بیماری پیشرونده است، به معنی اینکه علائم و نشانه ها به مرور زمان بدتر می‌شوند. هر چند که بیماری پارکینسون در نهایت منجر به ناتوانی می‌شود، بیماری اغلب به صورت آهسته پیشرفت‌می کند، و خیلی از افراد سال‌های زیادی از زندگی تولیدی را بعد از تشخیص دارند. از این گذشته، برعکس دیگر بیماری‌های عصبی وخیم، بیماری پارکینسون درمان پذیر است. این راه‌ها شامل درمان دارویی، کاشتن وسیله‌ای است که مغز را تحریک می‌کند و یا جراحی می‌باشد. اولین نشانه بیماری پارکینسون بسیار سخت قابل تشخیص است، زیرا می‌تواند بسیار ظریف مانند تکان نخوردن شانه‌ها زمان راه رفتن، لرزش ملایم در انگشتان دست و منمن کردن در سخن گفتن باشد. ممکن است فرد مبتلا دچار کاهش انرژی شود، احساس افسردگی یا مشکل در خواب داشته باشد و یا ممکن است استحمام فرد، غذا خوردن یا انجام امور معمولی و روزمره دیگر وی طولانی تر گردد. به طور کلی علائم و نشانه‌های بیماری پارکینسون را می‌توان به هشت دسته تقسیم کرد مانند (لرزشTremor ،حرکات آهستهbradykinesia، سفت شدن ماهیچه‌ها، به هم خوردن تعادل، از دست دادن حرکات اتوماتیک بدن، اختلال در صحبت کردن، اشکال در بلع، جنونdementia). تشخیص این بیماری به صورت پزشکی بسیار مشکل و هزینه بر است. با توجه به اینکه بیماری پارکینسون دومین بیماری رایج مغز و اعصاب و یکی از شایع‌ترین انواع بیماری در جوامع بشری است، هر روز محققان در پی این هستند تا راه حلی ساده و کم هزینه برای تشخیص زودهنگام این بیماری بیابند. در تشخیص بیماری پارکینسون از روش‌های مختلفی استفاده شده است. از جمله این روش‌ها می‌توان به سیستم‌های تشخیص گفتار اشاره نمود. از آنجا که اغلب این بیماری را توسط نشانه‌های صوتی بیماران مبتلا به PD مانند کاهش بلندی و وضوح صدا، اختلال در کیفیت صدا شناسایی می‌کنند، این روش کاربرد زیادی در تشخیص دقیق این بیماری دارد.

روش تحقیق

در این مقاله برای دستیابی به نتایج بهتر، تعداد مدل‌ها را با استفاده از زیرمجموعه داده‌های متفاوت توسعه دهیم و یا از شرایط مختلف در روش مدل‌سازی انتخابی استفاده کنیم ایده کلی در شکل(۱) نشان داده شده است:



شکل ۱. اندیشه کلی روش‌های ترکیبی

اگر تعدادی الگوریتم طبقه‌بند پایه موجود باشد، می‌توان با ترکیب نتایج آن‌هابه دقت بالاتری رسید. ایده روش‌های ترکیبی به این صورت است که مجموعه‌ای از طبقه‌بندها را با داده‌های آموزشی ایجاد کرده و میزان صحت را با انجام عملیات رای‌گیری بر روی

فصلنامه علمی تخصصی ایده های نو در علوم، مهندسی و فناوری ، دوره ۷، شماره ۱، بهار ۱۴۰۲، ص ۵۷-۵۰
ISSN: 2588-3984
http://www.Tajournals.com

نتایج آن‌ها بدست می‌آوریم. نحوه محاسبه خطای این روش با فرض داشتن N طبقه‌بندکه خطای هر یک P است از رابطه زیر محاسبه می‌شود[۱۲] :

$$P(error) = \sum_{k=\frac{N}{2}+1}^{N} \binom{N}{k} P^k (1-p)^{N-k} \qquad (۱)$$

در واقع مفهوم ترکیب این است که خروجی چندین مدل با هم مخلوط شده تا به تصمیم‌گیری بهتر برسند. روش‌های ترکیبی در بیشتر حالت‌ها نتایج بهتری نسبت به طبقه‌بند منفرد دارند. این الگوریتم بر پایه استخراج ویژگی می‌باشد و مجموعه آموزش را برای طبقه‌بند پایه ایجاد می‌کند. برخلاف روش‌های طبقه‌بندی معمول، داده‌های آموزشی فقط یک نسبت ازکل مجموعه داده‌ها نیست بلکه به طور تصادفی به K زیرمجموعه تقسیم می‌شود و سپس تبدیل آنالیز اجزای اصلی (PCA) بر هر زیرمجموعه اعمال می‌گردد. زیرمجموعه‌های انتخاب شده می‌توانند به صورت جدا ازهم و یامشترک انتخاب شوند ولی برای رسیدن به پراکندگی مناسب از زیر مجموعه‌های جدا از هم استفاده شده است. روش پیشنهادی با تغییراتی در روش جنگل تصادفی ایجاد شده است که الگوریتم آن به شرح زیر است:

فاز آموزش

Given
- X : the object in the training data set (an N × n matrix)
- Y : the labels of the training data set (an N × 1 matrix)
- L : the number of classifiers in the ensemble
- K : the number of subsets
- $\{w_1, w_2, \ldots, w_c\}$: the set of class labels

For i = 1...L
- Prepare the Feature Vector
  - Split **F** (the feature set) into K subsets
  - For j = 1...K
    - Let $X_{i,j}$ be the data set X for the features in $F_{i,j}$
    - Eliminate from $X_{i,j}$ a random subset of classes
    - Apply PCA on $X_{i,j}$ to obtain the feature vector $C_{i,j}$
  - Construct final feature vector R by merge the $C_{i,j}$
- Build classifier D_i using (R,Y) as the training set

فاز طبقه بندی

- Calculate maximum voting
  $\sum_{j=1}^{L} d_{j,k} = max_{m=1}^{c} \sum_{j=1}^{L} d_{j,m}$

شکل ۲. فاز آموزش

به منظور ایجاد داده‌های آموزشی برای طبقه‌بند پایه، مجموعه ویژگی‌ها را به صورت تصادفی به k مجموعه (kپارامتر الگوریتم است) تقسیم کرده و سپس به هرمجموعه تبدیل آنالیز اجزای اصلی(PCA) اعمال می‌کنیم. حال برای هر مجموعه، cضریب آنالیز اجزای اصلیPCA (cبه صورت تجربی تعیین می‌شود) را انتخاب می‌کنیم و سپس درانتها با ادغام این c ضریب، بردار ویژگی نهایی ایجاد می‌گردد.بردار ویژگی نهایی k×c مولفه دارد. بردار ویژگی ایجاد شده را با استفاده از طبقه‌بند ماشین بردار پشتیبان SVMآموزش داده و پس از Lبار تکرار الگوریتم (L به صورت تجربی تعیین می‌شود)، داده‌های آزمایشی را با L طبقه‌بند ایجاد شده آزمایش می‌کنیم.سپس با استفاده ازروش رای‌گیری ماکزیمم، خروجی نهایی راتعیین می‌کنیم.

یافته ها

دیتاست مورد بررسی در این مقاله، توسط ماکس لیتل محقق دانشگاه آکسفورد، جمع‌آوری شد که شامل فرکانس سیگنال‌های صحبت کردن افراد است. این فایل شامل ۲۳ ستون و ۱۹۷ ردیف شامل مشخصات بیماران و افراد غیر بیمار است. رکوردهای این فایل شامل صداهای ضبط شده ۳۱ نفر از افرادی بوده که ۲۳ نفر از آنها دارای بیماری پارکینسون است. هر ستون در این مجموعه برای یک معیار ارزیابی صدا تعیین شده است. وضعیت بیماران در ستون وضعیت با مقدار ۰ یا ۱ تعریف شده است. مقدار ۱ مربوط به یک بیمار پارکینسون و مقدار ۰ مربوط به یک فرد عادی است. در مجموعه داده اولیه برای هر بیمار ۶ رکورد تعریف شده که نام بیمار در ستون نام مشخص شده است. این پایگاه داده برای تحقیق در سال ۲۰۰۸ بر روی وب سایت تحقیقاتی UCI قرار گرفته است[۵]. جدول۱،مشخصات این ویژگی ها را نمایش می‌دهد.

جدول۱. ویژگی‌های موجود در مجموعه داده اولیه

ردیف	نام	نوع داده	توضیحات
۱	Name	String	مشخصه افراد مورد بررسی
۲	MDVP:Fo(Hz)	Real	میانگین فرکانس صوتی
۳	MDVP:Fhi(Hz)	Real	بالاترین مقدار فرکانس صوتی
۴	MDVP:Flo(Hz)	Real	کمترین مقدار فرکانس صوتی
۵	MDVP:Jitter(Abs), MDVP:Jitter(%) MDVP:PPQ , MDVP:RAP Jitter:DDP	Real	تغییرات مربوط به فرکانس صوتی
۶	MDVP:Shimmer , MDVP:Shimmer(dB) MDVP:APQ , Shimmer:APQ5 Shimmer:DDA	Real	معیارهای تغییرات مربوط به دامنه صوتی هر فرد
۷	NHR, HNR	Real	دو معیار سنجش وجود نویز در صدا
۸	Status	Real	وضعیت فرد ۱ برای بیمار پارکینسون

۰ برای فرد عادی			
دو معیار سنجش غیر خطی صوت	Real	RPDE , D2	۹
معیاری برای سنجش نویزهای تنفس در صدا	Real	DFA	۱۰
۳ معیار غیر خطی	Real	Spread1, Spread2, PPE	۱۱
۳ معیار غیر خطی	Real	Spread1, Spread2, PPE	۱۱

به منظور پیاده‌سازی الگوریتم‌های کلاسه‌بندی کننده روی نمونه‌های حاصل از ۱۹۷ مورد بررسی، از روش اعتبارسنجی متقابل[1] چند لایه با مقادیر K=3,5,10 استفاده شده است. در این روش داده‌ها به ۳، ۵ و ۱۰ قسمت مساوی تقسیم شده و در هر تکرار یکی از قسمت‌ها به عنوان داده تست و بقیه قسمت ها داده آموزشی در نظر گرفته می‌شوند. بنابراین الگوریتم کلاسه بندی سه، پنج یا ۱۰ بار اجرا می شوند. در نهایت مقدار میانگین دقت کلاسه بندی در این تکرارها به عنوان نتیجه نهایی الگوریتم در نظر گرفته می‌شود. در کلاسه‌بندی کننده درخت تصمیم از شاخص جینی برای تقسیم و ایجاد زیردرخت‌ها استفاده شده است. در کلاسه‌بندی کننده شبکه عصبی از الگوریتم شبکه عصبی پیش خور چند لایه استفاده شده است. الگوریتم مورد استفاده برای آموزش، الگوریتم لونبرگ-مارکوات است.. برای ارزیابی کلاسه‌بندی کننده‌ها معیارهای مختلفی از جمله ویژگی، حساسیت، دقت و ... وجود دارد. ماتریس اغتشاش شامل اطلاعاتی در خصوص نحوه کلاسه بندی رکوردها برای هر الگوریتم کلاسه‌بندی است. این اطلاعات شامل اطلاعات واقعی و اطلاعات پیش بینی بدست آمده از هر کلاسه‌بندی کننده است [۸]. در این بخش از معیار دقت برای تشخیص بیماری پارکینسون استفاده نمودیم. این معیار بر اساس فرمول زیر محاسبه می‌گردد.

جدول ۲. ارزیابی کلاسه بندی کننده ها با مقادیر مختلف

K-fold=3					
	Acc(%)	TN	FP	FN	TP
Neural Net	۹۳,۸۵	۴۳	۵	۷	۱۴۰
Decision Tree	۹۰,۷۷	۳۸	۸	۱۰	۱۳۹
Naïve Bayes	۹۰,۲۶	۳۱	۲	۱۷	۱۴۵
K-fold=5					
	Acc(%)	TN	FP	FN	TP
Neural Net	۹۳,۳۳	۴۳	۸	۵	۱۳۹
Decision Tree	۹۰,۲۶	۴۱	۱۲	۷	۱۳۵
Naïve Bayes	۸۹,۷۴	۳۲	۴	۱۶	۱۴۳
K-fold=10					

[1] K-fold Cross Validation

	Acc(%)	TN	FP	FN	TP
Neural Net	۹۳,۳۴	۴۰	۵	۸	۱۴۲
Decision Tree	۹۲,۸۴	۴۱	۷	۷	۱۴۰
Naïve Bayes	۹۱,۳۲	۳۳	۲	۱۵	۱۴۵

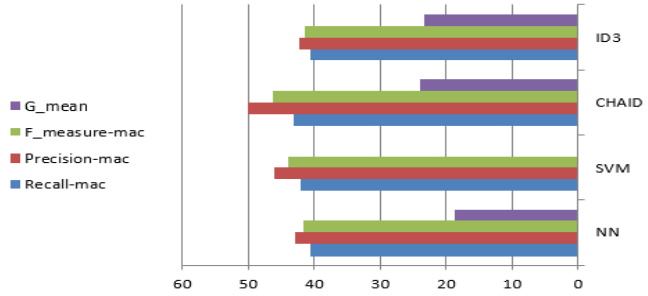
براساس جدول۲، شبکه عصبی در هر سه حالت، بیشترین دقت را داشته‌اند. اما مدت زمان محاسبه آن بسیار بالا است . مدت زمان محاسبه الگوریتم شبکه عصبی برای رکوردهای این پایگاه داده در حالت k-fold=10، ۱۶۲ ثانیه بوده است. بعد از شبکه عصبی، الگوریتم درخت تصمیم بهترین نتیجه را کسب کرده است. در این الگوریتم از شاخص جینی برای تقسیم ویژگی‌ها استفاده شده است. تعداد تشخیص های اشتباه الگوریتم بیزین از مقایسه با سایر الگوریتم‌ها کمتر بوده است. در حالت k-fold=3,10، این الگوریتم فقط ۲ مورد تشخیص اشتباه در بیماری پارکینسون داشته است (FP=2). اما در تشخیص افراد سالم، خطای الگوریتم بالا است (TN= 31,33). در جدول ۳ نتایج ادغام هر سه الگوریتم با الگوریتم AdaBoost آورده شده است. در روش Boosting رای هر کلاسه‌بندی کننده وزن مخصوص به خود را دارد. پس از اعمال وزن، هر کلاسی که بیشترین رای را بیاورد، برنده محسوب می‌شود.

جدول ۳. ادغام هر سه الگوریتم با الگوریتمAdaBoost

K-fold=3					
	Acc(%)	TN	FP	FN	TP
Neural Net	۹۰,۷۷	۳۹	۹	۹	۱۳۸
Decision Tree	۹۵,۹۰	۴۳	۳	۵	۱۴۴
Naïve Bayes	۸۹,۷۴	۳۴	۶	۱۴	۱۴۱
K-fold=5					
	Acc(%)	TN	FP	FN	TP
Neural Net	۹۳,۸۵	۴۰	۴	۸	۱۴۳
Decision Tree	۹۳,۳۳	۴۲	۷	۶	۱۴۰
Naïve Bayes	۸۸,۲۱	۳۵	۱۰	۱۳	۱۳۷
K-fold=10					
	Acc(%)	TN	FP	FN	TP
Neural Net	۹۱,۳۴	۴۲	۱۱	۶	۱۳۶
Decision Tree	۹۳,۲۹	۳۸	۳	۱۰	۱۴۴
Naïve Bayes	۸۷,۶۳	۳۴	۱۰	۱۴	۱۳۷

درخت تصمیم در حالت k-fold=3 بهترین نتیجه را بدست آورده است. تعداد تشخیص‌های اشتباه در این حالت در بین افرادی که بیماری پارکینسون داشته‌اند، نسبت به سایر الگوریتم‌ها بسیار کمتر بوده است (FP=3). مدت زمان اجرای الگوریتم شبکه عصبی در حالت k-fold=10 برابر با ۱۸۲۶ ثانیه بوده است. دقت الگوریتم درخت تصمیم با استفاده از AdaBoost در مقایسه با روش معمولی بهبود پیدا کرده است. تلفیق با الگوریتم AdaBoost باعث شده است که از بین تعداد بیماران پارکینسون، ۱۴۴ مورد تشخیص درست و تنها ۳ مورد تشخیص اشتباه محاسبه شود. درصد دقت در این روش با استفاده از درخت تصمیم به ۹۵,۹۰ درصد رسیده است. بعد از الگوریتم درخت تصمیم، شبکه عصبی بهترین نتیجه را در ادغام با AdaBoost داشته است.

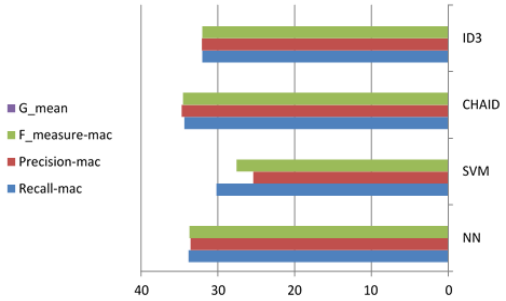
فصلنامه علمی تخصصی ایده های نو در علوم، مهندسی و فناوری ، دوره ۷، شماره ۱، بهار ۱۴۰۲، ص ۵۷-۵۰
ISSN: 2588-3984
http://www.Tajournals.com



	Recall-mac	Precision-mac	F_measure-mac	G_mean
NN	40.55	43.01	41.74	18.78
SVM	42.04	46.15	43.99	0.00
CHAID	43.04	50.07	46.28	24.07
ID3	40.63	42.23	41.40	23.35

شکل ۳. مقایسه الگوریتم‌های پایه (حاصل اعمال مدل روی مجموعه آزمون)

همان‌طور که شکل (۴) نشان می‌دهد، کارایی الگوریتم CHAID از سایر الگوریتم‌های شرکت کننده در این مقایسه بیش‌تر است.



	Recall-mac	Precision-mac	F_measure-mac	G_mean
NN	33.82	33.55	33.68	0.00
SVM	30.18	25.38	27.57	0.00
CHAID	34.37	34.72	34.54	0.00
ID3	32.02	32.06	32.03	0.00

شکل ۴. مقایسه الگوریتم‌های پایه (حاصل اعمال مدل روی مجموعه آموزش)

بحث و نتیجه‌گیری

در این مقاله با توجه به نتایج به دست آمده از مدل مشاهده شد که اغلب الگوریتم‌های مورد استفاده بالای ۸۰ درصد در درست بودن کلاسه‌بندی تشخیص بیماران مبتلا به پارکینسون و افراد سالم نقش داشته‌اند. همچنین الگوریتم های Adaboost و درخت تصمیم با درصد خطای ناچیز و قابل چشم‌پوشی موفق به ارائه بهترین تمیز این دو دسته از یکدیگر گشتند. الگوریتم پیشنهادیاین مقاله به ترکیب دسته بندهای داده کاوی پرداخته شد. ایده روش‌های ترکیبی به این صورت است که مجموعه‌ای از طبقه‌بند - کننده‌ها را با داده‌های آموزشی ایجاد کرده ومیزان صحت را با انجام عملیات رای‌گیری بر روی نتایج آن ها بدست می‌آوریم. در روش پیشنهادی داده‌ها را با استفاده از دسته بند ماشین بردار پشتیبانSVM آموزش داده سپس با استفاده از روش رای گیری

ماکزیمم خروجی نهایی را تعیین شد. نتایج به دست آمده از مدل پیشنهادی را با سایر مدل‌ها مقایسه نموده و نتایج نشان دهنده بهبود عملکرد این الگوریتم در تشخیص بیماران پارکینسون می‌باشد.

منابع

1. Kincade K. “Data mining: Digging for healthcare gold. Insurance & Technology”, 1998.
2. Fayyad U, Piatetsky-Shapiro G & Smyth P., “Knowledge discovery and data mining towards a unifying framework. Available” , 1996.
3. Koh HC & Tan G, “Data mining applications in healthcare. Journal of Healthcare InformationManagement”, ۲۰۰۵.
4. Han J, Kamber M & Pei J, “Data mining: Concepts and techniques. USA: Morgan Kaufmann Publishers Inc; ۲۰۲۱
5. Lee IN, Liao SC & Embrechts M, “Data mining techniques applied to medical information”, MedicalInformatics and the Internet in Medicine, ۲۰۰0.
6. Obenshain MK, “Application of data mining techniques to healthcare data”, Infection Control and Hospital Epidemiology, ۲۰۰۴.
7. Samad Soltani Heris T, Lagarizadeh M, Mahmoodvand Z & Zolnoori M, “Intelligent diagnosis ofasthma using machine learning algorithms. International Research Journal of Applied and Basic Sciences”, 20۲۰.
8. Liao SC & Lee IN, “Appropriate medical data categorization for data mining classification techniques”, Medical Informatics and the Internet in Medicine, 2022.
9. Maimon OZ, Rokach L. “Data Mining And Knowledge Discovery Handbook”. New York: Springer Science & Business; 2010.
10. Samad Soltani Heris T, Lagarizadeh M, Mahmoodvand Z & Zolnoori M, “Intelligent diagnosis ofasthma using machine learning algorithms. International Research Journal of Applied and Basic Sciences”, 201۹.