پیش بینی تقاضای دارو به کمک روشی ترکیبی مبتنی بر *KNN* و *SVR*
(چالش پیش بینی تقاضای دارو)

سعید جعفری زاده*[1]، فاطمه محمد پور[2] و** احمد رضا نقش نیلچی[3]

[1]دانشگاه اصفهان، saeid.jafarizadeh@gmail.com

[2]دانشگاه پاسارگاد، fatemehmohamadpoor91@gmail.com

[3]دانشگاه اصفهان، nilchi@eng.ui.ac.ir

*سرپرست تیم

**استاد راهنمای پروژه

چکیده – این چالش در مورد پیش بینی مقادیر مصرف دارو به صورت روزانه، و به تفکیک هر بیمارستان تا ۱۰ماه آتی با توجه به داده های سه سال قبل می باشد، روش ارائه شده در این تحقیق ترکیبی از روش *KNN* و *SVR* است . بر اساس نتایج حاصله از پیاده‌سازی روش‌های مختلف بر روی داده‌های ارائه شده در این چالش به این نتیجه رسیدیم که به دلیل کمبود ویژگی‌های مناسب(تنها یک ویژگی زمان و روز هفته دراختیار می‌باشد) وهمچنین ماهیت مسئله (پیش بینی برای یک دوره زمانی بسیار طولانی) اغلب مدل‌های پیش بینی از روش‌های سنتی و ساده گرفته تا روش‌های عمیق از جمله *LSTM* و *RNN* و *CNN* در اینجا عملکرد مناسبی را از خود نشان نمی‌دهند. بنابراین انتخاب روشی جهت استخراج ویژگی مناسب (جهت تغییر فضای ویژگی) در این مسئله بسیار حائز اهمیت است لذا نوآوری ارائه شده در این تحقیق تغییر فضای ویژگی و سپس مدل سازی بر اساس ویژگی جدید می‌باشد. (ذکر این نکته بسیار حائز اهمیت است که در اینجا از داده‌ای خارج از داده‌های ارائه شده استفاده نشده است)، نتایج نشان دهنده خطای پایین روش ارائه شده نسبت به روش های ذکر شده می باشد.

کلید واژه- KNN، SVR ، تقاضای دارو، پیش بینی ، LSTM ، CNN

۱- مقدمه

مصرف داروها در حوزه بهداشت و درمان اهمیت بسیاری دارد و پیش‌بینی مصرف داروها نیز از اهمیت بالایی برخوردار است. پیش‌بینی مصرف داروها به ما کمک می‌کند تا درمان‌ها را بهتر برنامه‌ریزی کنیم، عوارض جانبی را کاهش دهیم و هزینه‌ها را بهینه‌سازی کنیم.

در مورد پیش‌بینی مصرف داروها، می‌توان از رویکردهای متنوعی استفاده کرد. در زیر به برخی از روش‌ها و مسائل مرتبط با پیش‌بینی مصرف داروها اشاره می‌کنم:

۱. مسئله پیش‌بینی تقاضای دارو: در این مسئله، هدف پیش‌بینی تقاضای آینده برای یک دارو است. می‌توان از داده‌های گذشته مصرف دارو، اطلاعات پزشکی، مشخصات بیماران و سایر متغیرهای مرتبط استفاده کرد تا به یک مدل پیش‌بینی برای تقاضای آینده دارو برسیم. این مدل می‌تواند به متخصصان بهداشت و سازمان‌های تولید و توزیع دارو کمک کند تا توانایی تامین دارو را بهینه‌سازی کنند[1].

۲. مسئله پیش‌بینی مصرف دارو در جمعیت خاص: در این مسئله، هدف پیش‌بینی مصرف یک دارو در یک گروه خاص از جمعیت است. مثلاً می‌توانیم بررسی کنیم که چگونه مصرف یک دارو در افراد مبتلا به یک بیماری خاص (مانند دیابت) تغییر می‌کند. با استفاده از داده‌های بالینی و زیرمجموعه‌های مرتبط، می‌توانیم مدلی بسازیم که بتواند مصرف دارو را در این گروه خاص پیش‌بینی کند. این اطلاعات می‌تواند به پزشکان و محققان در تصمیم‌گیری‌های بالینی و تحقیقاتی کمک کند[2].

۳. مسئله پیش‌بینی عوارض جانبی دارو: در این مسئله، هدف پیش‌بینی عوارض جانبی یک دارو است. با استفاده از داده‌های بالینی و اطلاعات مربوط به عوارض جانبی، می‌توان یک مدل پیش‌بینی برای شناسایی و پیش‌بینی عوارض جانبی دارو ساخت. این مدل می‌تواند به پزشکان و سازمان‌های بهداشتی کمک کند تا در انتخاب و استفاده از داروها، ریسک‌های مرتبط با عوارض جانبی را در نظر بگیرند و تصمیمات بهتری داشته باشند[3].

۴. مسئله پیش‌بینی پاسخ به درمان: در این مسئله، هدف پیش‌بینی پاسخ بیماران به یک درمان خاص است. با استفاده از داده‌های بالینی، ویژگی‌های بیماران و اطلاعات مرتبط، می‌توانیم مدلی بسازیم که بتواند پاسخ بیماران به یک درمان را پیش‌بینی کند. این اطلاعات می‌تواند به پزشکان در انتخاب و تنظیم درمان‌ها و بهبود نتایج بالینی کمک کند[4-5].

در این تحقیق هدف " مسئله پیش‌بینی تقاضای دارو" در سطح بیمارستان های یک شهر است. در این مسئله تقاضای دارو هر روز ۱۲ بیمارستان مختلف به تفکیک برای یک سال به ما داده شده و از ما پیش بینی مصرف داروی ۱۰ ماه آینده این ۱۲ بیمارستان به تفکیک خواسته شده است. بدیهی است برای حل این مسئله راه حل های مختلفی وجود دارد . که به صورت کلی به دو دسته زیر تقسیم بندی می شوند:

- روش‌های کلاسیک: از جمله روش‌های ساده و متداول برای پیش‌بینی،

روش‌های خطی می‌باشند. در این روش‌ها، با استفاده از متغیرهای مستقل مانند تاریخ، فصل، مشخصات بیمارستان و سایر متغیرهای مرتبط، می‌توان یک مدل خطی را ساخت و با استفاده از آن، مصرف دارو را برای ۱۰ ماه آینده پیش‌بینی کرد. روش‌هایی مانند رگرسیون خطی ، ماشین بردار پشتیبان ، درخت تصمیم ، مدل های فازی و شبکه های عصبی ساده برای این منظور مورد استفاده قرار می‌گیرند.

- روش های مبتنی بر شبکه‌های عمیق: شبکه‌های عصبی عمیق می‌توانند روش هایی قدرتمند برای پیش‌بینی باشند. در حوزه یادگیری عمیق، شبکه‌های عصبی بازگشتی (RNNs) مانند LSTM و GRU و شبکه‌های عصبی پیچشی (CNNs) مورد استفاده قرار می‌گیرند. این شبکه‌ها قابلیت استخراج ویژگی‌های پیچیده از داده‌ها را دارند و می‌توانند در پیش‌بینی مصرف داروها و بسیاری از مسائل مرتبط با پیش بینی مورد استفاده قرار گیرند. با این حال روش های عمیق نیازمند داده‌های بزرگ و سیستم پردازشی بالا می باشند.

به طور کلی، برای انتخاب روش مناسب برای پیش‌بینی مصرف داروها، باید به ماهیت داده‌ها، تعداد متغیرها، حجم داده و هدف نهایی توجه کرد. در اینجا تنها داده های در دسترس، زمان و مقدار داروی مصرفی در هر روز است لذا حجم داده و تعداد ویژگی ها در وهله اول بسیار ناکافی به نظرمی رسد. لذا در این تحقیق روشی را برای غلبه بر این موضوع پیشنهاد می کنیم. در ادامه ابتدا به بیان روش های پیش زمینه (SVR و KNN) پرداخته و سپس به بیان روش استخراج ویژگی های مناسب برای حل چالش مسئله می پردازیم، در نهایت الگوریتم کلی ارائه شده برای حل مسئله شرح داده خواهد شد.

۲- روش پیشنهادی

به طور خلاصه در روش ارائه شده ابتدا ویژگی تاریخ را به روز در سال ، ماه و روز در ماه تقسیم بندی می کنیم تا سه ویژگی جدید حاصل شود و سپس به کمک الگوریتمی مشابه الگوریتم KNN ، این ویژگی ها را به ۱۰ ویژگی جدید از میان دادههای داده شده نگاشت کرده و فضای ویژگی جدیدی تولید میکنیم و به کمک این ویژگی ها یک مدل SVR خطی را آموزش می دهیم در نهایت و در مرحله تست برای هر داده جدید مشابه همین عمل را انجام داده و پس از محاسبه نگاشت ویژگی داده جدید از مدل آموزش دیده جهت پیش بینی مقدار خروجی داده تست مورد نظر استفاده می کنیم . برای درک راحت تر این مفهوم ابتدا الگوریتم SVR و KNN را شرح داده و سپس نحوه نگاشت ویژگی را به کمک الگوریتم KNN با ذکر جزییات شرح می دهیم.

۲-۱- الگوریتم SVR

الگوریتم SVR یا Support Vector Regression یک روش یادگیری ماشین برای پیش‌بینی مقادیر پیوسته است. این روش بر اساس ایده‌ی Support Vector Machine (SVM) که یک روش دسته‌بندی است، ساخته شده است. هدف این روش این است که یک تابع خطی یا غیرخطی را بیابد که با حداقل خطای ممکن(MSE)، مقادیر هدف را پیش‌بینی کند. برای این کار، این روش از یک مفهوم به نام ماشین بردار پشتیبان با حاشیه نرم (Soft Margin Support Vector Machine) استفاده می‌کند. در SVRهدف کاهش فاصله بردارهای پشتیبان از خط مارجین می‌باشد. این فاصله را می‌توان با استفاده از یک تابع به نام ماشین بردار پشتیبان با حاشیه نرم (Soft Margin Support Vector Machine) محاسبه کرد. این تابع به این معنی است که برخی از نقاط داده می‌توانند از حاشیه خارج شوند و خطایی را ایجاد کنند، اما این خطا با یک پارامتر به نام C کنترل می‌شود. هر چه مقدار C بزرگتر باشد، خطای بیشتری مجاز است و هر چه کوچکتر باشد، خطای کمتری مجاز است. این مفهوم به این معنی است که برخی از نقاط داده می‌توانند از حاشیه خارج شوند و خطایی را ایجاد کنند، اما این خطا با یک پارامتر به نام C کنترل می‌شود. هر چه مقدار C بزرگتر باشد، خطای بیشتری مجاز است و هر چه کوچکتر باشد، خطای کمتری مجاز است. این روش از یک تابع هسته (Kernel Function) نیز برای ایجاد یک فضای ویژگی جدید استفاده می‌کند که در آن داده‌ها به صورت خطی قابل جداسازی هستند. معمولا از توابع هسته‌ی گوسی (Gaussian Kernel)، چندجمله‌ای (Polynomial Kernel)، سیگموید (Sigmoid Kernel) و خطی (Linear Kernel) استفاده می‌شود در اینجا از تابع خطی به عنوان تابع کرنل استفاده شده است. برای یافتن تابع پیش‌بینی، این روش از یک تابع هزینه (Cost Function) به نام ε-insensitive loss استفاده می‌کند که در اینجا تابع هزینه . این تابع هزینه به این معنی است که اگر خطای پیش‌بینی کمتر از یک مقدار ثابت به نام ε باشد، هزینه صفر است و اگر بیشتر از آن باشد، هزینه برابر با اختلاف خطا و ε است. برای حل این مسئله، از یک روش به نام Lagrange Multipliers استفاده می‌شود که به دنبال بیشینه‌کردن حاشیه بین داده‌ها و تابع پیش‌بینی است. به طور کلی در نهایت، تابع پیش‌بینی به صورت زیر بدست می‌آید:

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b \qquad (۱)$$

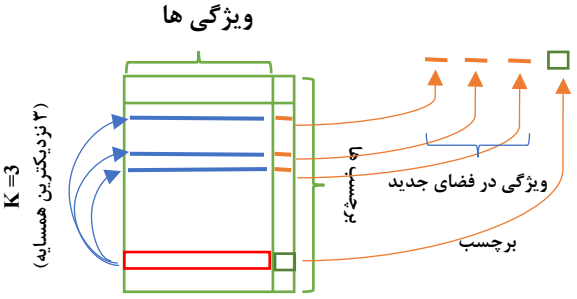
که در آن n تعداد داده‌ها، α_i و α_i^* ضرایب لاگرانژ، K تابع هسته(که در اینجا تابع خطی است) و b یک عدد ثابت است. برای یافتن این ضرا یب و عدد ثابت، از یک روش به نام Sequential Minimal Optimization (SMO) استفاده می‌شود که به صورت تکراری، دو ضریب را به روزرسانی می‌کند تا به شرایط KKT برسند. شرایط KKT شرایط لازم و کافی برای بهینه‌سازی یک تابع هزینه با محدودیت‌های خطی یا غیرخطی است.

۲-۲- الگوریتم KNN

الگوریتم KNN یا K-Nearest Neighbor یک روش یادگیری ماشین برای دسته‌بندی و پیش‌بینی مقادیر پیوسته است. این روش بر اساس شباهت نقاط داده با یکدیگر کار می‌کند. برای پیش‌بینی برچسب یا مقدار یک نقطه جدید، این روش از K نزدیک‌ترین همسایه آن در مجموعه داده آموزشی استفاده می‌کند. برای مثال، اگر K برابر ۳ باشد، برچسب یک نقطه جدید برابر با برچسب بیشترین تکرار در میان سه نزدیک‌ترین همسایه آن خواهد بود. این روش یک روش غیرپارامتریک است که به معنی این است که هیچ فرضی در مورد توزیع داده‌ها نمی‌کند. این روش می‌تواند با داده‌های عددی و رده‌ای کار کند و برای مسائل مختلفی در دسته‌بندی و پیش‌بینی کاربرد دارد. برای فاصله بین نقاط داده، از معیارهای مختلفی مانند فاصله اقلیدسی، فاصله منهتن، فاصله مینکوفسکی و تابع هسته استفاده می‌شود. برای یافتن نزدیک‌ترین همسایه‌ها، از الگوریتم‌های مختلفی مانند الگوریتم حریصانه، الگوریتم KD-Tree و الگوریتم Ball-Tree استفاده می‌شود.

۳-۲- نحوه نگاشت ویژگی ها به فضای ویژگی جدید به کمک الگوریتم KNN

در اینجا به دلیل ماهیت مسئله که رگرسیون می باشد از روش KNN برای برچسب دهی و دسته بندی استفاده نشده است بلکه تنها برای بدست آوردن K داده مشابه به نمونه در حال آموزش یا تست بهره گرفته ایم به این صورت که K نزدیکترین داده را از بین داده های آموزشی انتخاب می کنیم سپس برچسب این K داده را به عنوان ویژگی جدید (و بدون در نظر گرفتن ویژگی های قبلی) برای مرحله آموزش و یا تست مورد استفاده قرار میدهیم، به صورت نشان داده شده در شکل زیر:

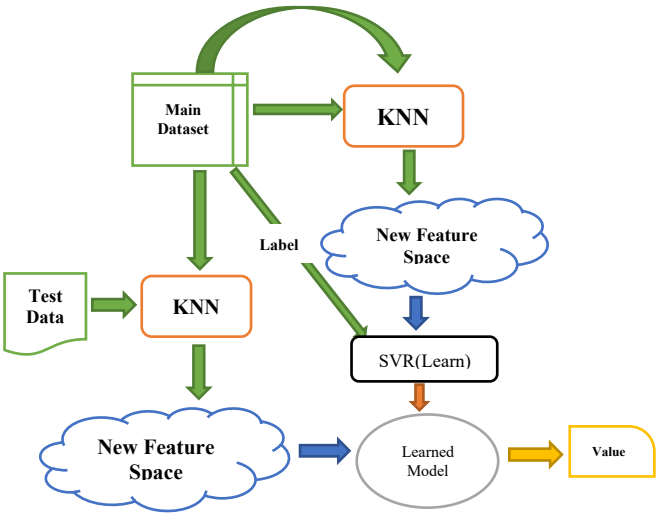


شکل ۱: نحوه تغییر فضای ویژگی با توجه به داده های موجود در مسئله پیش بینی تقاضای دارو

به عبارت دیگر در اینجا فضای ویژگی را از فضای زمانی به فضای برچسب ها نگاشت کرده ایم ، در ادامه می بینیم با توجه به نتایج حاصله در این فضا مدل هایی مانند بردار پشتیبان ماشین عملکرد بهتری را نسبت به روش های دیگر از خود نشان می دهند.

۴-۲- مدلسازی و آموزش

حال که روش تغییر فضای ویژگی ها شرح داده شد، در ادامه از روش SVR جهت مدلسازی و پیش بینی داده ها استفاده می کنیم، ابتدا هر یک از داده های دیتاست را به فضای جدید برده و سپس داده ها را بدون اعمال نرمال سازی به مدل SVR خطی می دهیم و مدل را با آن آموزش می دهیم ، در این مرحله برای هر یک از داده‌ها مقدار پیش بینی را از دیتا ست انتخاب می‌کنیم. به عبارت دیگر برچسب مربوطه را به ازای هر یک از داده‌ها از دیتا ست مورد نظر بیرون کشیده و مدل مورد نظر را آموزش می‌دهیم در مرحله تست نیز ابتدا ویژگی های مورد نظر را در فضای جدید از پایگاه دانش به کمک الگوریتم KNN داده شده بدست می آوریم و در نهایت مقدار پیش‌بینی نهایی را به کمک مدل آموزش داده شده در مرحله قبل محاسبه می کنیم ، شکل ۲ نشان دهنده نحوه عملکرد روش پیشنهادی مورد نظر و نحوه پیش بینی داده های جدید می باشد.



شکل ۲: تغییر فضای ویژگی از حالت ویژگی های مبتنی بر زمان به میژگی های مبتنی بر برچسب خروجی

به طوری کلی در این روش همانطور که در شکل بالا مشاهده می شود از داده های آموزشی هم در مرحله تست و هم در مرحله آموزش استفاده شده است ، در این روش بر خلاف روش های معمول مدلسازی (همانند SVR) پس از آموزش نیز نیازمند داده های دیتاست می باشیم و بر خلاف روش هایی مانند KNN در مرحله آموزش اقدام به ایجاد مدلی خطی از داده و تنظیم پارامتر های مدل مذکور نموده و در مرحله تست نیز از این مدل استفاده مینماییم!
در این روش پارامتر K برابر با ۱۰ در نظر گرفته شده و کرنل SVR نیز همانطور که گفته شد به صورت خطی در نظر گرفته شده است (نتایج نشان دهنده عملکرد بهتر کرنل خطی نسبت به کرنل های غیر خطی در این مسئله است.) همچنین مدل SVR مورد نظر در زمان آموزش به کمک روش اعتبار سنجی 10-fold بهینه می شود .

۳- نتیجه‌گیری

در این تحقیق علاوه بر روش پیشنهادی از روش هایی همانند LSTM دو گانه و یکتا به صورت ارائه شده در مرجع [6] و CNN (به صورت یک بعدی بر روی پنجره ای از داده ها به عنوان سری زمانی) نیز جهت حل مسئله استفاده شده است. اما نتایج نشان دهنده عملکرد ضعیف تر این الگوریتم ها در قبال روش ترکیبی ارائه شده می باشد . همچنین با توجه به تغییر فضای ویژگی ، ویژگی های تغییر یافته را نیز برای هر دو روش LSTM و CNN گفته شده بکار گرفته ایم اما کماکان نتایج حاصله نشان دهنده دقت بالاتر روش پیشنهادی می باشد بطوری که اختلاف خطای ۵ تا ۱۰ برابری نسبت به روش پیشنهادی در زمان اعتبار سنجی را شاهد بودیم. برای مثال جدول ۱ نشان دهنده بهترین نتایج حاصله برای پیش بینی داده های بیمارستان ۱ می باشد

جدول ۱ - مقایسه روش ارائه شده با مدل های مشابه برای بیمارستان ۱

روش / معیار	CNN	LSTM	SVR	Our Method
MAE	17.60	7.05	17.61	0.7601

در اینجا معیار ارزیابی میانگین مطلق خطا(MAE) در نظر گرفته شد است که به صورت فرمول ۳ محاسبه می شود.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i| \quad (3)$$

که در آن $\hat{y}$ مقدار پیش بینی شده و y مقدار واقعی میزان مصرف دارو در روز می باشد. به عنوان تحلیل نتایج حاصله می توان دلیل عدم موفقیت مدل های عمیق را در این مسئله به دلیل نداشتن داده کافی ، عدم غنای کافی فضای ویژگی ها، انتظار پیش بینی طولانی مدت (۱۰ ماه آینده) و مهمتر از همه تغییر شدید در داده های آموزشی بیان کرد به طوری که ممکن است در یک بیمارستان میزان مصرف دارو در یک روز رقم بالا (مثلا ۶۰ قلم دارو) و فردای آن روز تعداد بسیار پایین (مثلا ۱ یا دو قلم دارو) باشد (این امر هیچ وابستگی به ویژگی های ارائه شده مانند تعطیلی روز قبل یا بعد نیز ندارد و به کرات در داده دیتاست مشاهده می شود و از آنجایی که این امر به تعداد بسیار زیاد در دیتاست مشاهده شده نمیتوان این داده ها را به عنوان داده پرت در نظر گرفت!). طبیعتا با این اوصاف ، مدل هایی همانند LSTM توانایی پیش بینی تغییرات ناگهانی و پیش بینی دراز مدت را نداشته و لذا در اینجا بدیهی است که پیش بینی درستی را برای این تغییرات ارائه ندهند همچنین مدل هایی مانند CNN نیز بیش از حد به رفتار محلی سری زمانی وابستگی داشته و به تاریخ های مشابه همانند تاریخ های تکراری در سال یا ماه توجهی ندارد لذا روش پیشنهادی به نظر از لحاظ منطقی نیز بایستی به دلیل عدم وابستگی زیاد به حافظه کوتاه یا بلند مدت و همچنین عدم توجه زیاد به داده های محلی در آموزش نتیجه بهتری نسبت به مدل های مشابه داشته است.

سپاسگزاری

با تشکر از اساتید گرانقدر گروه هوش مصنوعی دانشگاه اصفهان که ما را در راستای انجام این پروژه یاری رسان بودند.

منابع

[۱] بیگلرخانی امین، عباسی رضوان، ثنایی محمدرضا. "پیش‌بینی میزان مصرف دارو در بیمارستان‌ها با استفاده از مدل شبکه حافظه طولانی کوتاه‌مدت. بیمارستان". ۱۴۰۱; ۲۱ (۴) :۲۲-۳۵.

[۲] مهاجر تبریزی, خوجه, درویش محمدی. "طراحی شبکهٔ توزیع دارو با استفاده از الگوریتم ژنتیک دوسطحی و کدینگ مبتنی بر اولویت (مطالعهٔ موردی: شرکت پخش سراسری آدوراطب)". مدیریت تولید و عملیات, ۲۰۲۲، ۱۳(۳), ۷۵-۴۷.

[۳] سندسی, گلشن, صائبی, هاشمی گلپایگانی. "یک روش پیش‌بینی پیوند مبتنی بر همسایه برای شبکه دوبخشی" فصلنامه فناوری اطلاعات و ارتباطات ایران, ۴۷(۴۷), ۱۷۵.

[۴] رستمی. پیش‌بینی کننده‌های بالینی پاسخ به درمان تحریک مکرر مغناطیسی فراجمجمه ای (rTMS) در بیماران مبتلا به اختلال افسردگی. فصل‌نامه پژوهش‌های کاربردی روانشناختی, ۲۰۱۷, ۸.۲: ۱۸۱-۲۰۰.

[۵] گلبابائی سروش, برهانی خاطره برهانی حامد. پیش‌بینی تصمیم‌گیری اخلاقی کادر درمان در شرایط غیرقطعی بر اساس مواجهه با مرگ، احتمال ابتلا و رضایت شغلی: نقش میانجی‌گری اضطراب. پژوهش های روانشناسی اجتماعی ۲۰۲۳, ۱۳(۴۹), ۱۴-۱.

[6] GAJAMANNAGE, Kelum; PARK, Yonggi; JAYATHILAKE, Dilhani I. “Real-time forecasting of time series in financial markets using sequentially trained dual-LSTMs”. Expert Systems with Applications, 2023, 223: 119879.