دوره ۱۳، شماره بهار و تابستان ۱۳۹۹

مجله فناوری اطلاعات در طراحی مهندسی
Information Technology in Engineering Design
http://ited.sinaweb.net

مدل مبتنی بر تنسور برای خوشه‌بندی داده‌های سلامت در راستای پزشکی دقیق

لیلا برادران سرخابی [۱] فرهاد سلیمانیان قره چیق * [۲] جعفر شهام فر [۳]

(۱) گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران

(۲) گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران *

(۳) گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران، گروه پزشکی اجتماعی، واحد تبریز، دانشگاه تبریز، تبریز، ایران

(تاریخ دریافت : ۱۳۹۹/۱۱/۰۵ تاریخ پذیرش : ۱۴۰۰/۴/۲۰)

چکیده

پزشکی دقیق یکی از انواع مدل‌های نوین پزشکی است که در صدد ارائه روش درمان متناسب با ویژگی‌های بیمار می‌باشد. برای ارائه روش درمانی یا پیشگرانه مناسب بیمار، ابتدا بیماران به زیرگروه‌هایی با بیشترین تشابه، تقسیم می‌شوند. دستیابی به پزشکی دقیق، نیازمند خوشه‌بندی صحیح و دقیق اطلاعات بیمار است، لذا استفاده از روش‌های داده‌کاوی که کارایی چشمگیری در خوشه‌بندی دارند انتخابی هوشمندانه به نظر می‌رسد. داده‌های حوزه سلامت، چند–وجهی و چند–بعدی با روابط پیچیده هستند و برای کاوش آنها نیاز به ساختاری داریم که قابلیت مدیریت چنین داده‌هایی را داشته باشد. ساختار ذاتی تنسورها باعث توانایی آنها در مواجهه با داده‌های چند–وجهی با ابعاد زیاد می‌گردد. در این مقاله مدلی مبتنی بر تنسور، در راستای خوشه‌بندی دقیق بیماران ارائه شده است. برای دستیابی به مدل نهایی فرایندی سیستماتیک معرفی گردیده که روند کامل انتقال داده‌های ورودی خام به محیط تنسور را شامل می‌شود. در این فرایند وجوه اصلی داده شناسایی شده و پس از پیش پردازش ابعاد تنسور نهایی را تشکیل می‌دهند. سپس تنسور تجزیه شده و خوشه‌ها استخراج می‌گردند. در راستای ارزیابی مدل پیشنهادی، سه منبع داده واقعی توسط مدل خوشه‌بندی شده و با نتایج حاصل از سه روش داده‌کاوی مقایسه شده‌اند. دقت و صحت خوشه‌بندی توسط مدل پیشنهادی به ترتیب ۲۱٪ و ۳۳٪ بیشتر از سایر روش‌ها بوده و زمان کاوش نیز علی‌رغم هزینه سربار آماده سازی داده، ۹٪ کاهش داشته است. همچنین اساس منطقی تنسور باعث می‌شود نتایج قابل اطمینان‌تر باشند و سریع بودن مدل‌های مبتنی بر تنسور باعث تحمل حجم بالای داده‌های پزشکی می‌شود.

کلید واژه: داده کاوی، تنسور کاوی، داده‌های پزشکی، پزشکی دقیق، تجزیه تنسور

*عهده‌دار مکاتبات:

فرهاد سلیمانیان قره چیق

نشانی: گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران

پست الکترونیکی: bonab.farhad@gmail.com تلفن ۰۹۱۴۱۷۶۴۴۲۷

۱ – مقدمه

در پزشکی نوین، باور بر این است که روش درمان، مانند لباس بدون سایز که برای همه اندازه باشد نیست و هسته پزشکی نوین در واقع تشخیص شیوه درمانی مناسب برای هر فرد است. هر چندکه گسترش کاربرد چنین مدل‌هایی می‌تواند مرزهای علم پزشکی جابجا کند، ولی با علم و فناوری‌های فعلی، چنین اقدامی برای تمامی بیماری‌ها و بیماران امکان-پذیر نیست. پزشکی دقیق یکی از مدل‌های پزشکی نوین است که در آن برای غلبه بر این چالش، بیماران خوشه‌بندی می‌شوند و بیمارانی که دارای ویژگی‌های خاص مشترکی هستند در یک گروه قرار می‌گیرند و سیاست تشخیص و درمان مشترکی برای کل اعضای گروه اتخاذ می‌شود [۱]. این مدل که پتانسیل بالایی در ارتقا کیفیت درمان و کاهش هزینه‌ها دارد دارای دو رکن اصلی است. یکی از این ارکان ژنتیک است ولی ژنتیک همه ماجرا نیست. پزشکی دقیق به عوامل محیطی و متغیرهای بسیاری نیز وابسته است [۲]. برای مثال پزشک می‌تواند بر اساس رژیم غذایی منحصر به شخص، یا باکتری‌های موجود در سیستم گوارش و یا بررسی عوامل خون فرد برای او روش درمانی مناسبی پیش بگیرد. بررسی زوایای مختلف مرتبط با سلامت نیازمند یکپارچه-سازی و تحلیل داده‌های متنوعی از منابع گوناگون است که همین موضوع باعث می‌شود علم داده دومین رکن اساسی پزشکی دقیق گردد. داده‌های حوزه سلامت منبع بسیار ارزشمندی در راستای تحقق پزشکی دقیق هستند. این داده‌ها که در ابتدا صرفاً جهت مکانیزاسیون پرونده‌های پزشکی ذخیره می‌شدند [۳]، هم اکنون به لطف بروز شاخه‌های نوین پزشکی به بحثی داغ در علم داده و پزشکی تبدیل شده و مورد توجه متخصصان هر دو حوزه قرار گرفته‌اند [۴]. داده-های پزشکی چند–وجهی، چند–رابطه‌ای و دارای تعداد زیادی بعد هستند. بهره‌گیری از دانش پنهان موجود در این داده‌ها نیازمند ابزار و ساختارهای مناسب است و با توجه به این ویژگی‌ها، استخراج اطلاعات قابل اعتماد و دقیق از آنها آسان به نظر نمی‌رسد.

تعداد زیاد ابعاد و وجوه همچنین وجود روابط متعدد پیچیده، از عمده‌ترین چالش‌های مطرح در داده‌های پزشکی هستند که باعث ناکارآمدی بسیاری از روش‌های کاوش (مانند خوشه‌بندی و دسته‌بندی) می‌شود [۵]. ساختار ذاتی و منطقی تنسورها، آنها را به انتخابی هوشمندانه برای کاوش و نمایش داده‌های پزشکی تبدیل می‌کند چرا که تنسور قابلیت مدیریت و تحمل داده‌های چند–وجهی و چند–بعدی را دارد. روش-های تجزیه تنسور راهکاری عملی برای خوشه‌بندی نمودن سریع و دقیق داده‌های چند–بعدی هستند که خوشه‌بندی مورد نیاز پزشکی دقیق را میسر می‌سازند و به دلیل پایه ریاضی تنسورها بسیار قابل اطمینان می‌باشند [۶–۷]. همچنین قابلیت تحلیل همزمان ابعاد، نه تنها سرعت کاوش را بسیار بالا می‌برد بلکه امکان استخراج روابط پیچیده پنهان مابین وجوه و ابعاد مختلف را نیز تسهیل می‌نماید [۸]. هرچند که محیط تنسور بسیار مناسب برای کاوش داده‌های پزشکی به نظر می‌رسد، استفاده این ساختار نیز به نوبه خود چالش-هایی به همراه دارد. برای مثال تعداد بالای ابعاد هزینه عملیات روی تنسور را افزایش می‌دهد و همچنین تجزیه تنسور و نمایش نتایج را نیز مشکل می‌نماید [۹]. چنانکه از تعداد ابعاد کمی برای کاوش داده‌ها استفاده شود، نتایجی بسیار فشرده و غیر قابل تفکیک و تفسیر آسان حاصل می‌گردد و از طرفی نیر استفاده از تعداد ابعاد زیاد هزینه زیادی دارد، بنابراین‌مدل تنسور مناسب، باید موازنه‌ای مابین نتایج تفسیرپذیر و هزینه کاوش برقرار نماید [۱۰]. یکی دیگر از چالش‌های اصلی در این حوزه، انتقال داده‌های خام پزشکی به مدل تنسور است

[۱۱].این داده‌ها نه تنها مناسب استفاده در محیط تنسور نیستند بلکه جزو کثیف‌ترین داده‌های موجود در جهان می‌باشند [۱۲]. برای حل این مسئله نیاز به وجود فرایندی است که در وهله اول داده‌ها را پاکسازی نماید تا نتایج، قابل اطمینان باشند و در وهله دوم آنها را برای تطابق با محیط تنسور ویرایش آماده نماید.

در این مقاله مدلی مبتنی بر ساختار تنسور برای کاوش داده‌های پزشکی پیشنهاد شده و مراحل تشکیل مدل نیز ارائه گردیده است. هدف اصلی مدل، افزایش دقت، صحت و سرعت خوشه‌بندی در راستای تحقق کاربرد گسترده پزشکی دقیق است. اینکار با بررسی همزمان ابعاد مختلف داده انجام می‌شود. برای کاهش هزینه‌های عملیاتی، ابتدا طی الگوریتمی کاهش بعد انجام می‌گیرد و ابعاد و وجوه اصلی داده شناسایی می‌شوند. این وجوه، ابعاد مدل را شکل داده و سپس با استفاده از روش تجزیه تنسور، خوشه‌های مورد نظر به دست آمده‌اند. برای ارزیابی مدل پیشنهادی، نتایج حاصل از کاوش روی سه منبع داده واقعی با نتایج سه روش متداول داده‌کاوی مقایسه شده است. همچنین برای سنجش میزان هزینه و دقت مدل پیشنهادی، داده‌ها توسط دو مدل مبتنی بر تنسور دیگر نیز خوشه‌بندی شده‌اند. در راستای افزایش قابلیت اطمینان، ابتدا داده‌های خام پالایش شده و سپس کاوش روی آنها انجام گرفته است.

۲ - کارهای پیشین

داده‌کاوی یکی از روش‌های محبوب برای افزایش کیفیت خدمات و کاهش هزینه‌ها در پزشکی است. در دهه اخیر تحقیقات بسیاری در خصوص کاربرد داده‌کاوی در علم پزشکی انجام شده است [۱۵-۱۳]. شاخه‌های نوین علم پزشکی نیز کاملاً وابسته به علم داده بوده و روز به روز استفاده از روش‌های داده‌کاوی در این حوزه رایج‌تر می‌شود [۱۷-۱۶]. پزشکی دقیق نیز یکی از همین شاخه‌هاست که محققان زیادی در تلاش برای بهره‌گیری از داده‌کاوی در راستای گسترش عملیاتی شدن آن هستند [۱-۲]. استفاده از تنسور در داده‌کاوی [۱۸] و تحقق پزشکی نوین بحثی داغ در هر دو حوزه داده و پزشکی است [۱۹]. مروری بر روش‌های مبتنی بر تنسور و کاربرد آنها در تحلیل داده‌های پزشکی دارد. کاربرد تنسور در کاوش داده‌های پزشکی نتایج بسیار چشمگیری داشته است [۲۰]. از تنسور به عنوان زیرساخت اصلی برای تحلیل روابط مابین بیمار، تشخیص و درمان استفاده نموده و توسط تجزیه تنسور قادر به آشکارسازی اتوماتیک ابعاد پنهان این سه وجه شده است [۶]. با مرور بر آخرین روش‌های مبتنی بر تنسور و موفقیت آنها در افزایش دقت داده‌کاوی، تنسور را به عنوان ساختاری مناسب برای استفاده در پزشکی دقیق معرفی نموده است. همچنین در این مقاله، تنسور به عنوان ساختاری با قابلیت تحمل به‌روز رسانی‌های مکرر داده‌های پزشکی عنوان شده است [۲۱]. در از تنسور در راستای پزشکی دقیق بیماران قلبی با ویژگی‌های خاص استفاده شده و طبق نتایج به دست آمده تفسیرپذیری در این روش به دلیل یکپارچه سازی و کاهش ابعاد در تنسور بسیار بیشتر بوده است [۲۲]. از روش تجزیه تنسور برای خوشه‌بندی داده‌های پزشکی استفاده نموده که طبق نتایج بسیار دقیق‌تر از روش‌های کلاسیک بوده است [۲۳] و [۲۴]. از تجزیه تنسور برای استخراج دانش پنهان در داده‌های بیمارستانی و تبدیل آنها به مفاهیم قابل تفسیر پزشکی استفاده نموده‌اند [۲۵]. پا را فراتر نهاده و تنسور را به عنوان ساختاری قابل انتقال به زیرساخت‌های بسط پذیر معرفی نموده و در

محیط ابری کاوش و نمایش داده‌های پزشکی را با نتایج بسیار مقبولی انجام داده است.

۳ - روش پیشنهادی

در این بخش به شرح مدل پیشنهادی برای نمایش و کاوش دقیق داده‌های سلامت خواهیم پرداخت. داده‌های سلامت داده‌هایی چند-وجهی و چند-بعدی هستند. بدین معنی که داده دارای وجوه مختلفی بوده و هر وجه نیز ابعاد مختلفی را شامل می‌شود. ارتباطات پیچیده‌ای که مابین این ابعاد وجود دارند در مدل‌های خطی و دو-بعدی قابل تشخیص نیستند. همچنین الگوها و متغیرهای پنهان فراوانی در میان این وجوه نهفته است که تشخیص وبهره‌برداری از آنها می‌تواند نتایج کاوش را بسیار دقیق‌تر و کاراتر نماید. در این راستا، در این مقاله به آنالیز داده‌های واقعی پزشکی پرداخته و پس از تعیین وجوه و ابعاد آنها مدلی مناسب این گونه داده ارائه نمودیم. داده‌های سلامت گستره بسیار وسیعی دارند، از داده‌های دیجیتال ایجاد شده توسط کاربران در دستگاه‌های هوشمند تا عکس‌های پیشرفته رادیولوژی [۲۶]. لیکن قسمت اعظم داده-های پزشکی را رکوردهای سلامت الکترونیک[1] تشکیل می-دهند [۱۰]. لذا در این مقاله مدل پیشنهادی بر اساس رکوردهای سلامت الکترونیک خواهد بود، لیکن تصور می-شود که مدل پیشنهادی، قابل بسط در سایر انواع داده‌های سلامت نیز باشد.

۳ – ۱ توصیف داده‌های پزشکی

داده‌های پزشکی نمونه بارزی از کلان داده هستند و به طبع آن تمام خصوصیات کلان داده را دارا خواهند بود [۲۷]. علاوه بر آن داده‌های پزشکی برخی خصوصیات اختصاصی دارند که باعث شده که کاوش آنها بسیار چالش برانگیزتر از سایر داده‌ها باشد [۲۸]. برای ارائه مدل مناسب با داده‌های هر حوزه لازم است که آنها به دقت بررسی شوند، لذا در این تحقیق به بررسی دقیق داده‌های در دسترس و عمومی پرداختیم [۲۹-۳۱]. جدول ۱، ویژگی‌های برجسته داده‌های پزشکی، دلایل عمده وجود این ویژگی‌ها و تاثیر آنها در کاوش داده‌ها را نشان می‌دهد. حجم بالا، وجود نویز بسیار زیاد، ناهمگنی، تنوع داده و وجود تعداد زیاد وجه و بعد در داده‌های پزشکی، باعث شده نمایش و کاوش آنها کاری آسان نباشد. علاوه بر آن تحقیقات نشان می‌دهد اکثر ابزارهای ذخیره‌سازی، داده‌های پزشکی را به صورت طولی ذخیره‌می-نمایند [۳۲]. این نوع ذخیره‌سازی و رفتار با داده‌های پزشکی، مشکل کاوش و نمایش آنها را دو چندان می‌سازد. وجود وجوه و ابعاد زیاد در داده‌های پزشکی از دلایل اصلی بروز مشکلات بسیاری در کاوش و ارائه نتایج است. در این مقاله مدلی ارائه می‌شود که قابلیت مدیریت تعداد بالای وجوه و ابعاد داده را داشته باشد.

[1] Electronic Health Records (EHRs)

جدول ۱) ویژگی‌های داده‌های پزشکی، دلایل بروز این ویژگی‌ها و تاثیر آنها در کاوش داده

ویژگی	دلایل وجود ویژگی	چالش‌ها
ناهمگنی داده‌ها	عدم وجود ساختار و فرمت استاندارد	مشکل بودن کاوش و نمایش داده‌ها عدم امکان/مشکل بودن تشخیص افزونگی و داده‌های یکسان بروز ناسازگاری
حجم بالای نویز	اشتباه عمدی/غیرعمدی در ورود داده عدم وجود اطلاعات کافی در زمان ثبت داده	از مسیر اصلی خارج نمودن الگوریتم‌های کاوش نتایج کاوش غیرقابل اعتماد و اشتباه
حجم بالای داده	گسترش ابزارهای هوشمند کنترل و ثبت داده ثبت داده‌های پزشکی به صورت الکترونیک رشد سریع و به روز رسانی فراوان داده‌ها	پرهزینه شدن کاوش
ابعاد بسیار زیاد	ثبت متا داده و داده‌های اصلی زیاد لزوم بررسی ابعاد مختلف در تشخیص و درمان	پرهزینه شدن کاوش لزوم استفاده از الگوریتم‌های نوین پیچیده برای نمایش و کاوش
وجوه زیاد	ارتباط سلامت با سایر وجوه زندگی ارتباط تنگاتنگ حوزه سلامت با سایر حوزه‌ها	مشکل بودن تشخیص و نمایش ارتباطات بیم وجوه مختلف
تنوع داده‌ها	جمع آوری داده‌ها از منابع مختلف فرمت و ساختار متفاوت در داده‌های سلامت	بروز ناسازگاری پرهزینه بودن کاوش

۳ – ۲ پزشکی دقیق و داده‌کاوی

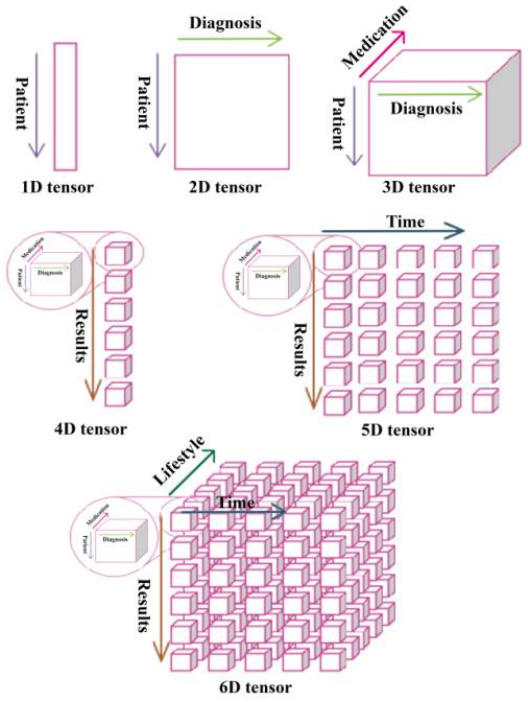
پزشکی دقیق یکی از شاخه‌های پزشکی نوین است که هدف آن ترکیب و تحلیل داده‌هایی است که در طول زمان از یک فرد یا گروه مشخص گردآوری می‌شود تا بتواند از آنها برای تشخیص/پیشگیری درست بیماری و ارائه روش درمانی مناسب افراد استفاده کند. چرا که فرضیه اصلی در پزشکی دقیق این است که روش درمان مانند لباس بدون سایز نیست که مناسب همه افراد باشد بلکه بر خلاف پزشکی سنتی در پزشکی دقیق ویژگی‌های بیمار نیز به اندازه بیماری یا حتی بیشتر از آن مورد بررسی قرار می‌گیرد. داده‌های مورد بررسی در پزشکی دقیق، شامل داده‌های ژنتیکی، محیطی، سبک زندگی و هر آنچه که بتواند به وضعیت سلامت فرد وابسته

باشد، هستند [۳۳]. اولین قدم در فرایند پزشکی دقیق خوشه-بندی بیماران بر اساس شباهت‌شان در ویژگی‌های خاص است که بر اساس آن سیاست مشترکی برای تشخیص و درمان اعضای یک گروه اتخاذ می‌شود پرا که با علم حاضر امکان ارائه روش درمانی و داروی انحصاری برای افراد وجود ندارد . وظیفه خوشه‌بندی دقیق بیماران یکی از اصلی‌ترین چالش‌ها در پزشکی دقیق است، چرا که بیماران باید در ساختاری با وجوه و ابعاد زیاد یکدیگر مقایسه شده و مشابه-ترین افراد در یک گروه قرار گیرند. داده‌کاوی یک انتخاب هوشمند برای اینکار است چرا که تحلیل داده‌های پزشکی با خصوصیات فوق‌الذکر، با روش‌های کلاسیک اگر غیرممکن هم نباشد بسیار مشکل خواهد بود. علم داده بعد از علم ژنومیکس موثرترین عامل در مسیر دستیابی به پزشکی دقیق است. در انجام کاوش کارا و قابل اعتماد فاکتورهایی مانند زیرساخت، مدل، روش و نمایش مناسب دخیل هستند. مدل مناسب داده‌های سلامت باید اول اینکه، قابلیت مدیریت داده-هایی با وجوه و ابعاد بالا و حجیم را داشته باشد. دوم اینکه، کاوش دقیق و آشکارسازی روابط پنهان بین داده‌های سلامت نیازمند یکپارچه‌سازی آنهاست که متعاقباً منجر به بروز ساختاری پیچیده‌تر با تعداد بیشتری وجه و بعد خواهد شد و مدل باید توانایی مدیریت آنها را داشته باشد. سوم اینکه، داده‌های مرتبط با سلامت، عموماً از منابع ناهمگن مختلفی چون رکوردهای الکترونیک سلامت، دستگاه‌های هوشمند و غیره گردآوری می‌شوندکه از فرمت تا ادبیاتشان با هم تفاوت دارد لذا لازمه کاوش یکپارچه آنها نیازمند ابزاری مناسب است که بتواند این داده‌های ارزشمند را به صورت منسجم و با قابلیت نمایش و تفسیر بالا مدل کند. ارائه چنین مدلی می-توانند مسیر ارائه سلامت همگانی را هموار نمایند.

۳ – ۳ تنسور و داده‌های پزشکی

تنسورها ابزاری قدرتمند و کارا برای مدل کردن، نمایش و کاوش انواع داده‌های ناهمگن چندوجهی با ابعاد زیاد هستند[۱۹]. در واقع تنسورها حالت بسط داده شده ابعاد ماتریس، یا همان آرایه های چند–بعدی هستند که به هریک از ابعاد آن وجه گفته می‌شود [۹]. به نظر می‌رسد تنسور انتخابی هوشمند برای مدل نمودن داده‌های سلامت باشد چرا که ساختار ذاتی تنسورها به ما امکان می‌دهد ضمن حفظ ساختار اصلی داده، آنها را با حالتی تفسیرپذیرتر و قابل فهم‌تر مدل نماییم. کاوش داده نیز در تنسور با استفاده از روش‌های تجزیه تنسور امکان‌پذیر است. تجزیه تنسورها راهکاری بسیار سودمند برای استخراج اطلاعات نهفته در داده‌های چند–وجهی می‌باشد و در داده‌کاوی نتایج بسیار موفق و کارایی داشته اند [۳۴–۲۴] همین امر باعث شده که تنسورها از بحث‌های داغ سال‌های اخیر در حوزه داده‌کاوی بوده و در کاربردهای گسترده‌ای در انواع زمینه‌ها از تحلیل داده‌های شبکه‌های اجتماعی تا تحلیل داده‌های پزشکی حضور مورد توجه واقع شوند. داده‌های پزشکی از جمله داده‌های چند–وجهی با ابعاد بسیار زیاد هستند که سرعت افزایش حجم آنها قابل توجه است. این داده‌ها را منابع مختلفی چون سیستم‌های اطلاعات بیمارستانی، اپلیکیشن‌های همراه و سنسورهای مختلف تولید میکنند و همانطور که اشاره شد، ویژگی‌های ذاتی تنسورها بسیار مناسب برای چنین داده‌هایی است. از زیرساخت تنسور می‌توان برای مدل نمودن و متعاقباً آنالیز داده‌های پزشکی استفاده نموده و روابط پیچیده مابین وجوه مختلف را شناسایی کرد چرا که ساختار تنسور امکان تفسیر پذیری و نمایش بهتر داده‌های چند–وجهی را فراهم می-نماید [۱۱]. پس از مدل نمودن داده در ساختار تنسور می‌توان از روش‌های تجزیه تنسور برای کاوش استفاده نمود. لذا

تجزیه تنسور می‌تواند جهت خوشه‌بندی داده‌های پزشکی و قدم نهادن در مسیر پزشکی دقیق، مورد استفاده قرار گیرد. شکل ۱، نحوه نسبت دادن وجوه مختلف داده‌های پزشکی در تنسورهایی با درجه‌های مختلف را نشان می‌دهد. همانطور که مشاهده می‌شود تنسور با درجه یک، همان بردار و تنسور با درجه دو، همان ماتریس است که تحلیل و کاوش آنها بسیار متداول بوده و روش‌های مقبولی برای انجام آنها وجود دارد، مشکل از جایی آغاز می‌شود که بخواهیم برای کاوش دقیق‌تر، ابعاد بیشتری از داده را مدل نماییم.



شکل ۱) نحوه مدل نمودن داده‌های پزشکی در تنسور با ابعاد مختلف

۳ – ۴ مدل پیشنهادی

اولین قدم در ارائه مدل‌های مبتنی بر تنسور مشخص نمودن ساختار تنسور یعنی تعداد ابعاد آن است. تعداد ابعاد نقش بسزایی در کارایی عملیات تنسور دارد. هر چه ابعاد تنسور بیشتر باشد عملیات کاوش هزینه پر هزینه تر و هر چه ابعاد کمتر باشد نتایج کاوش کلی‌تر و دارای دقت کمتری خواهد بود [۳۵]. در این تحقیق برای تعیین تعداد ابعاد مدل پیشنهادی، در ابتدا داده‌های چندین منبع مختلف به دقت بررسی گردیدند. پس از آنالیز داده‌ها مشخص گردید که‌علی‌رغم تعداد بالای ابعاد، گروه‌های محدودی در داده‌های پزشکی وجود دارند. بدین معنی که هر بعدی وابسته به یکی از گروه‌ها هستند. گروه‌ها همان موجودیت‌های اصلی در حوزه هستند که از ابعاد مختلف برای ثبت و نمایش مقادیر ویژگی‌های آنها استفاده می‌شود. برای مثال مشخصات فردی و سبک زندگی فرد همه از ابعاد مرتبط با موجودیت بیمار هستند. بیمار یکی از موجودیت‌های اصلی در داده‌های پزشکی است. برای تعیین موجودیت‌های اصلی یا همان گروه‌ها که متعاقباً وجوه تنسور را تشکیل خواهند داد از روش

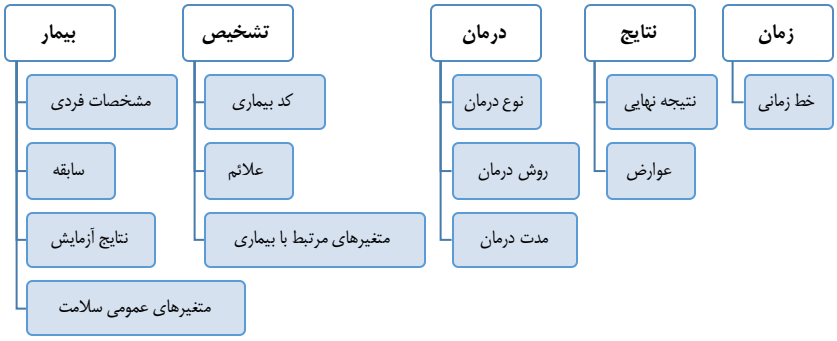
دستی استفاده شد و داده‌ها گروه‌بندی شدند. برای گروه‌بندی داده ابتدا کل ابعاد داده بررسی گردیده و موجودیت‌های اصلی در داده‌ها شناسایی شدند. سپس هر یک از ابعاد به یکی از موجودیت‌ها نسبت داده شد. برا تعیین موجودیت‌ها از الگوریتم پایین به بالا استفاده شده و در ابتدا برای هر یک از ابعاد از داده یک سرگروه مجزا در نظر گرفته می‌شود، در هر مرحله از الگوریتم، سرگروه‌های مرتبط با یک موجودیت با یکدیگر ترکیب می‌شوند [۳۶] و الگوریتم تا زمانی که نتوانیم سرگروه‌هایی را با هم ترکیب کنیم، ادامه می‌یابد. شکل ۲، نتایج خوشه‌بندی و برخی از وجوه اصلی مرتبط با هر سرگروه را نشان می‌دهد. هر یک از این سرگروه‌ها یکی از ابعاد مدل تنسور را تشکیل خواهند داد. جدول ۲، اطلاعات عمومی در مورد منابع مورد بررسی را نشان می‌دهد.

جدول ۲) توصیف منابع مورد بررسی

منبع	تعداد رکورد	تعداد بعد	تعداد رکورد ناقص	نوع ابعاد	نویز
۱	۸۰۰	۴۰۱	۴۷	ناهمگن	دارد
۲	۲۰۰۸	۱۶۶	۲۴۵	ناهمگن	دارد
۳	۶۵۰۷	۹۲	۱۱۸۴	ناهمگن	ندارد

در طول آنالیز داده‌ها به متغیر غیرمستقیمی برخوردیم که در عمل، سایر وجوه از این متغیر متاثر هستند. این متغیر، متغیر زمان است که توالی داده‌های ثبت شده در وجوه مختلف را نشان می‌دهد. به عبارت دیگر تمام داده ثبت شده در یک بعد خاص در زمان مشخصی قابل تعریف و معتبر است و چنانکه تاثیر متغیر زمان را در نظر نگیریم روند تشخیص و درمان و همچنین نتایج حاصل از آن ناممکن می‌شود. زمان، بعد پنهانی از داده‌های سلامت است که نقش بسیار مهمی در سابقه پزشکی بیمار، طول درمان و دوران پسا درمان دارد، لذا وجه زمان نیز به سرگروه‌ها اضافه شده است. لازم به ذکر است متاسفانه در وجه نتایج، تعداد اطلاعات ناقص بسیار زیاد بود و این موضوع به معنی عدم ثبت نتیجه میانی یا پایانی یک روش درمان برای یک بیماری و بیمار خاص است. در این مقاله برای کامل نموده اطلاعات ناقص، از الگوریتم کامل کردن با بیشترین پراکندگی[۲] [۳۷] با بهره‌گیری از داده‌هایی که روش درمانی مشابهی با داده ناقصی که قرار است تکمیل نماییم داشتند، استفاده شده است [۳۸].

[2] Imputation with mode



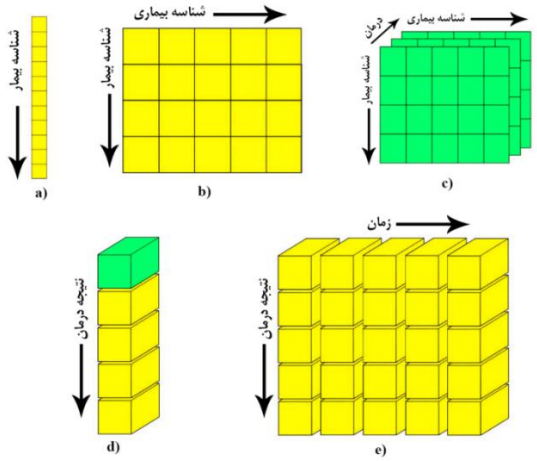
شکل ۲) گروه‌بندی وجوه مختلف داده‌های سلامت

پس از به دست آوردن وجوه تنسور، می‌توان داده را در آن مدل نمود. از آنجاییکه پنج گروه داریم در نتیجه مدل نهایی پنج وجهی خواهد بود. جدول ۳، مقادیر مورد استفاده برای ابعاد در وجوه مختلف را نشان می‌دهد. حال برای تشکیل تنسور نهایی نیازمند نگاشت داده‌های موجود به مدل هستیم.

مراحل نگاشت داده به تنسور پنج بعدی در شکل ۳ نشان داده شده است. بررسی و تحلیل همزمان این پنج بعد اصلی از داده‌های سلامت می‌تواند دقت نتایج را بسیار بالا برده و روابط پیچیده مابین آنها را آشکار سازد.

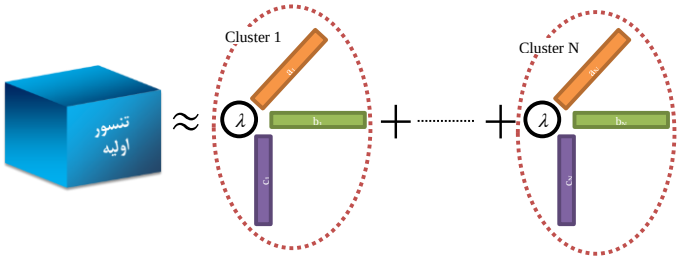
جدول ۳) وجوه و ابعاد مدل تنسور

وجه	بعد	شرح
بیمار	کد شناسایی منحصر به فرد بیمار	استفاده از این کد انحصاری برای ردیابی یک بیمار الزامی است.
بیماری	کد ICD 10 مربوط به بیماری	در منابع بومی اطلاعات این بعد بسیار ناقص است و داده‌ها با نظر دو کارشناس تکمیل گردید.
روش درمان	روش‌های کلی درمان به صورت طبقه بندی شده	روش کلی مورد استفاده در در درمان بیماری بدون در نظر گرفتن داروهای و روش مورد استفاده در این بعد استفاده شد.
نتیجه	بولین (مثبت / منفی)	نتیجه درمان به صورت بولی با حالت‌های مثبت و منفی
زمان	بازه‌های زمانی یک ماهه	این بعد حاوی بازه زمانی تاثیر روش درمان روی بیمار است



شکل ۳) مراحل تشکیل ابعاد مدل تنسور

برای کاوش داده‌ها در مدل تنسور از یکی از روش‌های رایج تجزیه تنسور به نام CP[3] استفاده خواهیم نمود که تنسور را به صورت مجموعی از ضرب خارجی چند بردار، نشان می‌دهد [۸]. نتیجه ضرب خارجی بردارها تنسوری با رتبه یک است. هر کدام از این تنسورهای رتبه یک، نمایانگر یک خوشه هستند. شکل ۴، مفهوم تجزیه تنسور با روش CP را نشان می‌دهد.



شکل ۴) تجزیه تنسور با روش CP

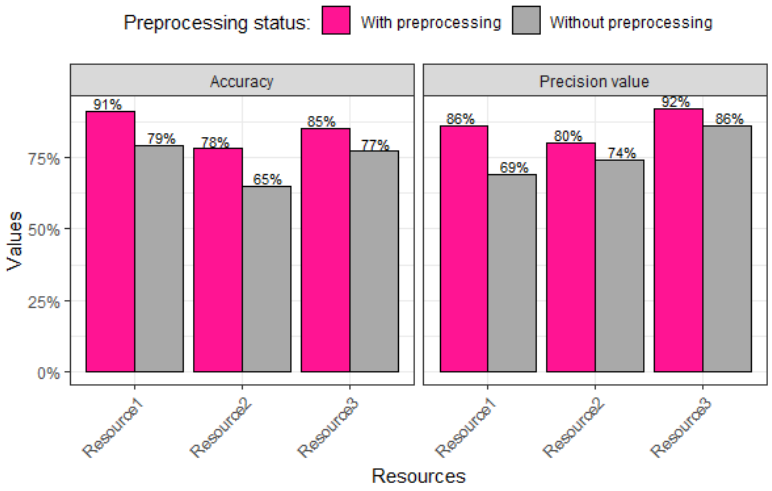
خوشه‌های حاصل از تجزیه تنسور نمایانگر زیرگروهی از بیماران هستند که شرایط آنها از لحاظ چهار بعد بیماری، روش درمان، نتیجه درمان و ترتیب زمانی بیشترین مشابهت را به یکدیگر دارند [۳۹]. در راستای افزایش دقت روش، فرایند پاکسازی روی داده‌های خام قبل از ورود به محیط تنسور اعمال گردید که نتایج نشان می‌دهد این فرایند تاثیر بسزایی در قابل اعتماد بودن نتایج دارد. فرایند پاکسازی اعمال شده برای مدل پیشنهادی شامل تشخیص داده‌های غیرمعمول [۴۰]، تشخیص داده تکراری [۴۱]، تکمیل داده‌های ناقص [۴۲]، جداسازی داده‌های عددی [۴۳] و خوشه‌بندی داده‌های متنی می‌باشد [۴۴]. به طور معمول فرایند پیش پردازش در دو مرحله انجام می‌گیرد. در مرحله اول داده پاکسازی شد و در مرحله دوم آماده انتقال به محیط مقصد (تنسور) می‌شود. هرچند که روش‌های بیشتری برای پاکسازی داده وجود دارد [۴۵]، در این تحقیق برای کاهش هزینه سربار فرایند

[3] CANDECOMP/PARAFAC

پاکسازی از چهار روش اصلی نامبرده شده استفاده گردیده است.

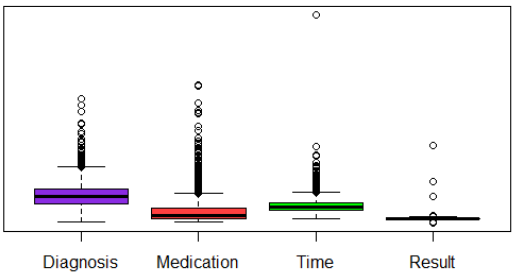
۴ – یافته‌ها

در این تحقیق از ساختار تنسور برای ارائه مدلی برای کاوش داده‌های سلامت در راستای پزشکی دقیق استفاده شده است. برای ارائه مدل به تحلیل دقیق داده‌های واقعی در حوزه سلامت پرداختیم. با توجه به تعداد بسیار زیاد ابعاد و وجوه در داده‌های پزشکی، جهت تسهیل روند کاوش و نمایش نتایج، ابتدا ابعاد را با الگوریتم پایین به بالا ترکیب نمودیم. این الگوریتم روی تمام منابع داده در دسترس اعمال شد و در نهایت در تمام منابع به چهار موجودیت اصلی مشترک رسیدیم. با اضافه نمودن زمان به عنوان یک موجودیت غیرمستقیم که بسیاری از ویژگی‌ها متاثر از آن هستند، مدل نهایی ما شامل پنج بعد گردید. پس از تعیین ابعاد تنسور نهایی برای انتقال داده‌های خام به محیط تنسور فرایند پیش‌پردازش روی آنها اعمال شد. این فرایند شامل تشخیص داده‌های غیرمعمول، تشخیص داده‌های تکراری، تکمیل داده‌های ناقص، جداسازی داده‌های عددی و گروهبندی داده‌های متنی بود. اعمال فرایند پیش‌پردازش منجر به حذف برخی از داده-های نویزدار گردید. شکل ۵ نتایج کاوش را در دوحالت نشان می‌دهد: با اعمال فرایند پیش‌پردازش و بدون اعمال فرایند پیش‌پردازش. همانگونه که مشاهده می‌شود پاکسازی داده‌های کثیف، تاثیر چشمگیری در افزایش دقت و صحت نتایج به خصوص در منابعی که تعداد داده ناقص آنها زیاد است، دارد.



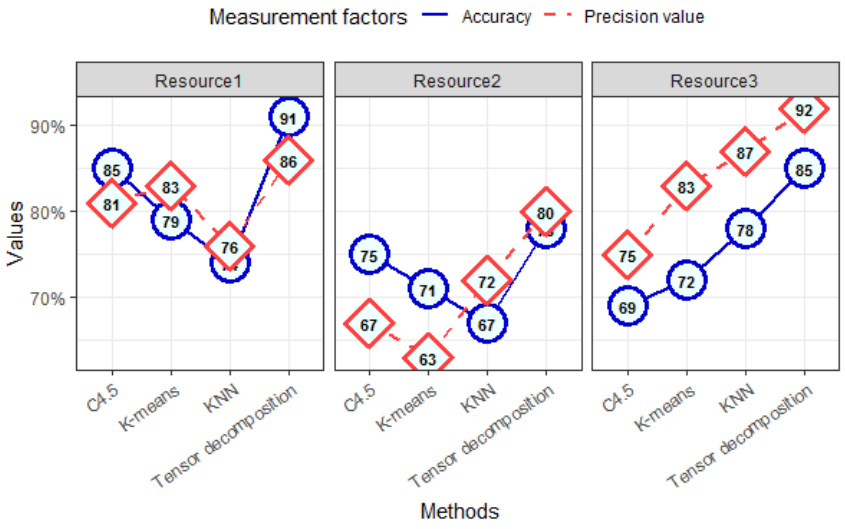
شکل ۵) تاثیر پیش پردازش در دقت و صحت نتایج

شکل ۶، نتایج عملیات تشخیص داده‌های غیرمعمول را با استفاده از روش تصویرسازی نشان می‌دهد که برای یافتن آنها از بسته ggplot2 زبان R استفاده شده است.



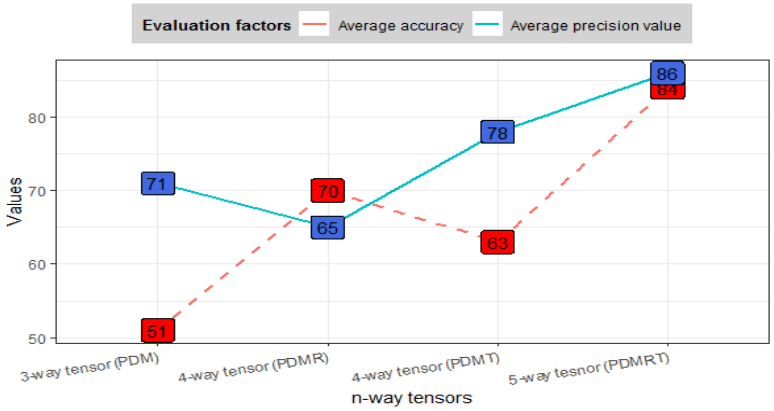
شکل ۶) تشخیص داده‌های غیرمعمول موجود در چهار بعد بیماری، روش درمان بیماری، نتیجه درمان و زمان

پس از پالایش داده‌ها، ۵٪ از داده‌های هر منبع به صورت تصادفی به عنوان داده تست انتخاب شد که برای ارزیابی دقت و صحت نتایج، این داده‌ها توسط متخصص حوزه بررسی گردید. داده‌ها پس از انتقال به ساختار تنسور با روش CP تجزیه شده و به خوشه‌هایی تقسیم شدند. هر کدام از این خوشه‌ها بیانگر داده‌هایی بودند که در تمام پنج بعد اصلی تنسور، بیشترین تشابه را با یکدیگر داشتند. گروه‌ها به ترتیب شامل بیماران با وضعیت پر خطر، خطر معمولی، کم خطر و بدون خطر می‌باشند. در راستای ارزیابی نتایج مدل پیشنهادی، سه روش داده-کاوی رایج برای جداسازی داده را روی داده‌ها اعمال نمودیم. شکل ۷، نتایج به‌دست آمده از چهار روش داده-کاوی اعمال شده روی سه منبع داده را نشان می‌دهد که معیارهای ارزیابی صحت و دقت توسط روابط زیر محاسبه شده‌اند:



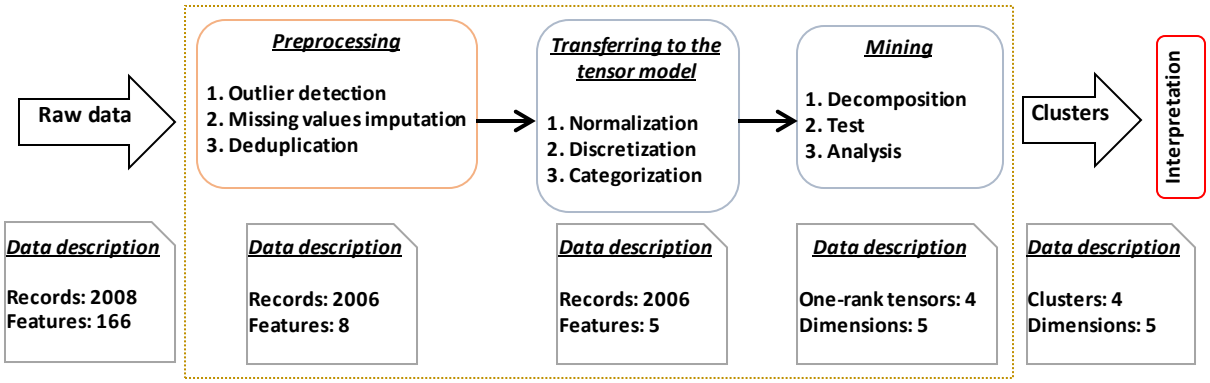
شکل ۷) میزان صحت و دقت چهار روش در تقسیم داده‌های سه منبع مختلف

از آنجاییکه نتیجه تجزیه تنسور منحصر بفرد نمی‌باشد، در راستای همگرایی نتایج، تجزیه تنسور پنج بار تکرار گردید و در هر تکرار موقعیت ابعاد تغییر یافت و نتیجه تجزیه بهینه با بررسی میزان دقت تعیین شد. همچنین برای سنجش میزان تاثیر ابعاد چهارم و پنجم (نتیجه درمان و زمان)، تنسور سه بعدی (بیمار، بیماری و درمان) با افزایش جداگانه ابعاد نیز تحلیل شد. نتایج این تحلیل در شکل ۸ قابل مشاهده است. طبق نتایج به دست آمده، بعد زمان و نتیجه درمان به ترتیب باعث افزایش ۱۲ و ۱۹ درصدی دقت کاوش می‌شوند. همانگونه که در شکل ۸ مشاهده می‌شود دقت مدل پیشنهادی ۳۳٪ بیشتر از مدل‌های تنسور سه بعدی می‌باشد.

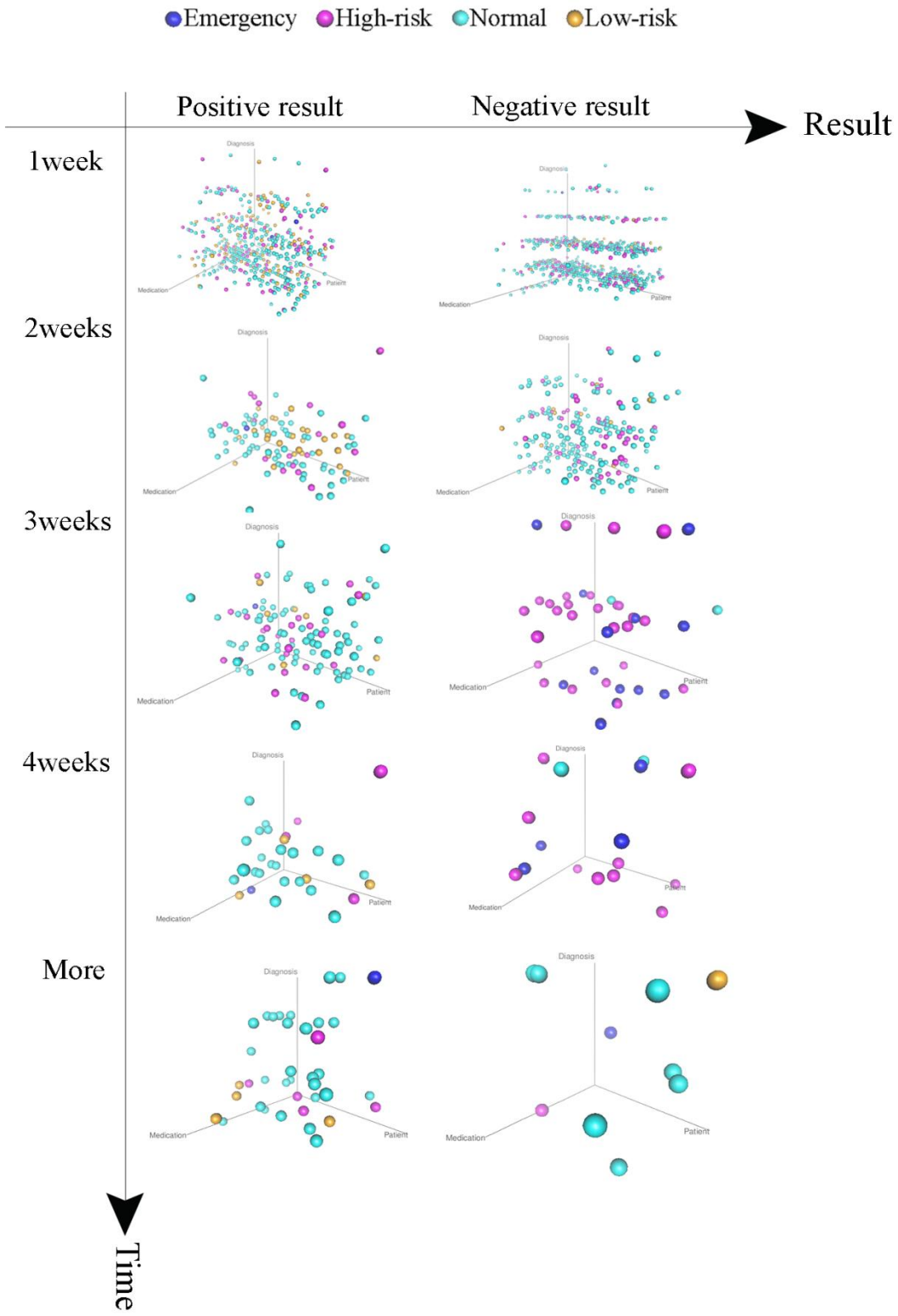


شکل ۸) صحت و دقت کاوش در تنسورهای چند-بعدی (P:patient, D:diagnosis, M:medication, R:Result, T:time)

هزینه زمانی کاوش در مدل پیشنهادی ۱۲٪ بیشتر از مدل سه بعدی بوده و هزینه سربار فرایند پیش پردازش به طور میانگین ۹٪ می‌باشد که در شکل ۹، نشان داده شده است. شکل ۱۰، عملیات هر مرحله از روند کاوش و تاثیر آن روی یکی از منابع داده را نشان می‌دهد. شکل ۱۱، نتایج کاوش توسط روش‌های مختلف روی همین منبع داده را نشان می‌دهد.



شکل ۹) مراحل مختلف فرایند کاوش و تاثیر هر کدام از آنها روی داده



شکل ۱۰) نتایج کاوش روی یکی از منابع داده

۵ – بحث

در تمام شاخه‌های پزشکی نوین علاوه بر بیماری، ویژگی‌های بیمار و عوامل مرتبط با او نیز مورد بررسی قرار می‌گیرد. فرضیه اصلی در علم نوین پزشکی بر این است که روش درمانی که برای شخصی نتیجه مثبت داده است لزوماً برای شخص دیگری موثر نخواهد بود. پزشکی دقیق یکی از این شاخه‌هاست که هدف اصلی در آن تشخیص تفاوت‌ها و شباهت‌های افراد و خوشه‌بندی دقیق آنهاست. این عمل در نهایت منجر به به افزایش کیفیت درمان، کاهش زمان تشخیص و کاهش هزینه‌ها خواهد شد. چنین خوشه‌بندی دقیقی علاوه بر تحلیل ژنتیکی، نیازمند بررسی همه جانبه داده‌های پزشکی است. این داده‌های ارزشمند چند–بعدی و چند–وجهی با روابط پیچیده هستند. هر اندازه که اعضای یک گروه از جهات مختلف مشابه به یکدیگر باشند، روش درمان ارائه شده کاراتر خواهد بود. لذا برای افزایش دقت خوشه‌بندی، نیازمند بررسی ابعاد و وجوه مختلف داده‌های پزشکی در محیطی یکپارچه هستیم. افزایش ابعاد، افزایش هزینه کاوش را در پی داشته و ساختارهای کلاسیک مانند ماتریس پاسخگوی نیازهای روش‌های داده‌کاوی در محیط چند–وجهی نیستند.

برای کاوش چنین داده‌هایی تنسور به دلیل ساختار منطقی ذاتی خود بسیار مناسب است. ابعاد مختلف داده می‌توانند در محیط تنسور یکپارچه شده و با استفاده از روش‌های تجزیه تنسور تفکیک گردند. در این مقاله با تحلیل داده‌های واقعی، مدلی مبتنی بر تنسور برای کاوش دقیق داده‌های پزشکی ارائه شده است. همچنین فرایند کامل ساخت مدل از پاکسازی داده‌های خام تا تجزیه تنسور شرح داده شده و برای ارزیابی مدل پیشنهادی، سه منبع داده واقعی توسط چهار روش (تجزیه تنسور، KNN، Kmeans و C4.5) تقسیم‌بندی شده‌اند. معیارهای ارزیابی دقت و صحت نتایج است. همچنین جهت بررسی زمان سربار فرایند ساخت مدل زمان کاوش نیز با روش‌های رایج دیگر مورد مقایسه قرار گرفته است. در مدل پیشنهادی، پس از تحلیل داده پنج وجه کلی دخیل در کارایی درمان، شناسایی شده و ابعاد اصلی مدل را تشکیل داده‌اند. این وجوه شامل: بیمار، بیماری، روش درمان، نتایج درمان و زمان هستند. لیکن برای بررسی کارایی مدل‌هایی با ابعاد کمتر و سنجش میزان تاثیر هر بعد در نتایج نهایی، مدل‌های سه–بعدی (بیمار، بیماری و روش درمان) و چهار–بعدی (بیمار، بیماری، روش درمان، نتیجه درمان/زمان) نیز مورد بررسی قرار گرفته است.

یک الی پنج درصد از داده‌های هر منبع به صورت تصادفی، به عنوان داده تست انتخاب شده و توسط دو کارشناس بررسی گردیده است. طبق نتایج به دست آمده، تحلیل همزمان ابعاد اصلی رکوردهای سلامت الکترونیک، دقت و صحت نتایج را نسبت به روش‌های کلاسیک به طور چشمگیری افزایش داده است. همچنین علی رغم زمان سربار فرایند تشکیل مدل، زمان کلی کاوش در مدل پیشنهادی نسبت به سایر روش‌ها تغییر چندانی ندارد. یکپارچه نمودن پنج بعد اصلی باعث افزایش قابل توجه دقت و صحت نسبت به مدل‌های سه–بعدی و چهار–بعدی شده است. نتایج نشان می‌دهد کاوش داده‌های پزشکی با مدل پیشنهادی نه تنها منجر به خوشه‌بندی دقیق و صحیح می‌شود بلکه به دلیل پایه ریاضی ساختار اصلی بکار رفته در مدل (تنسور) بسیار قابل اطمینان نیز می‌باشد.

۶ – نتیجه‌گیری

پزشکی دقیق در راستای عملی نمودن پزشکی متناسب با بیمار و برای رفع محدودیت‌های امکانات و تکنولوژی فعلی، روش تحلیل گروهی را جایگزین بررسی فردی نموده است. اینکار مستلزم خوشه‌بندی دقیق بیماران بر مبنای تشابه‌ها و تفاوت‌های آنها از ابعاد مختلف و با تکیه بر داده‌های پزشکی است. از آنجاییکه داده‌های حوزه سلامت داده‌هایی چند-وجهی، چند-رابطه‌ای و دارای ابعاد (صفات) بسیار زیاد هستند و ساختار ذاتی تنسور قابلیت تحمل این ویژگی‌ها را دارد، در این مقاله مدلی مبتنی بر تنسور برای کاوش این داده-ها در محیطی منسجم ارائه شده است. مدل پیشنهادی قابلیت انسجام ابعاد و وجوه مختلف داده را در یک محیط فراهم کرده و بررسی ابعاد در محیط یکپارچه به آشکارسازی روابط پنهان پیچیده مابین آنها کمک می‌کند. طبق نتایج حاصل از ارزیابی‌ها، دقت و صحت خوشه‌بندی در این مدل نسبت به روش‌های کلاسیک داده‌کاوی قابل توجه است و علی رغم فرایند پالایش و انتقال داده به مدل، هزینه زمانی سربار ندارد. با توجه به اینکه فرایند ساخت مدل بر اساس ویژگی‌های داده‌های پزشکی طراحی شده و ساختار ذاتی و منطقی تنسور باعث دستیابی سریع به نتایج قابل اطمینان با دقت و صحت بالا می‌شود، مدل پیشنهادی می‌تواند مسیر رسیدن به پزشکی دقیق را هموارتر نماید. لیکن بهره‌گیری از ساختار تنسور در راستای کاوش داده‌های پزشکی، هنوز در ابتدای این مسیر قرار دارد و برای رسیدن به نتایج مطلوب راه طولانی در پیش است. ارائه مدل‌هایی برای بررسی ابعاد بیشتر داده‌های پزشکی، مدیریت رشد و تغییر داده‌های پزشکی توسط مدل-های مقیاس‌پذیر و ایجاد زمینه‌های همکاری مدل‌های مبتنی بر تنسور و روش‌های نوین یادگیری ماشین، مانند یادگیری عمیق، بحث‌هایی هستند که در ادامه این مسیر می‌توانند وجود داشته باشند.

مراجع

۱. Huang BE, Mulyasasmita W, Rajagopal G. The path from big data to precision medicine. Expert Review of Precision Medicine and Drug Development. 2016;1(2):129-43. 10.1080/23808993.2016.1157686

۲. Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S, et al. From Big Data to Precision Medicine. Front Med. 2019;6(34). https://doi.org/10.3389/fmed.2019.00034

۳. Pashazadeh A, Navimipour NJ. Big data handling mechanisms in the healthcare applications: A comprehensive and systematic literature review .J Biomed Inform. 2018;82:47-62. https://doi.org/10.1016/j.jbi.2018.03.014

۴. Chawla NV, Davis DA. Bringing big data to personalized healthcare: a patient-centered framework. Journal of general internal medicine. 2013;39(55):660-5. 10.1007/s11606-013-2455۸-

۵. Ross MK, Wei W, Ohno-Machado L. "Big Data" and the Electronic Health Record. Yearb Med Inform. 2014;9(1):97-104. https://doi.org/10.15265/IY-2014-0003

۶. Yuan L, Fei W, Peter S. Tensor factorization toward precision medicine. Brief Bioinformatics. 2 ۴-۵۱۱:(۳)۱۸;۰۱۶ https://doi.org/10.1093/bib/bbw026

۷. Vitali F, Marini S, Pala D, Demartini A, Montoli S, Zambelli A, et al. Patient similarity by joint matrix trifactorization to identify subgroups in acute myeloid leukemia. JAMIA Open. 2018;1(1):75-86 . /۱۰.۱۰۹۳jamiaopen/ooy008

۸. Rabanser S, Shchur O, Günnemann S. Introduction to Tensor Decompositions and their Applications in Machine Learning. ArXiv. 2017 .

۹. Kolda TG, Bader BW. Tensor Decompositions and Applications. SIAM REVIEW. 2009;51(3):455-500 .

۱۰. Kruse CSA-Ohoo, Goswamy R, Raval Y, Marawi S. Challenges and Opportunities of Big Data in Health Care: A Systematic Review. Journal of Medical Internet Research, Medical Informatics. 2016;4(4):e38. 10.2196/medinform.5359

۱۱. ur Rehman MH, Liew CS ,Abbas A, Jayaraman PP, Wah TY, Khan SU. Big Data Reduction Methods: A Survey. Data Science and Engineering. 2016;1(4):265-84. 10.1007/s41019-016-0022-0

۱۲. Raghupathi W, Raghupathi V. Big data analytics in healthcare: Promise and potential. Health Inf Sci Syst. 2014;2(3):10. https://doi.org/10.1186/2047-2501-2-3

۱۳. Lan K, Wang D-t, Fong S, Liu L-s, Wong KKL, Dey N. A Survey of Data Mining and Deep Learning in Bioinformatics. Journal of Medical Systems. 2018;42(8):139-67. 10.1007/s10916-018-1003-9

۱۴. Sun J, Reddy C, editors. Big Data analytics for healthcare. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining 2013; Chicago, Illinois, USA. 10.1145/2487575.2506178

۱۵. Ahmad P, Qamar S, Afser Rizvi SQ. Techniques of Data Mining In Healthcare: A Review. Int J Comput Appl. 2015;120(15):38-50 .

۱۶. Beaulieu-Jones BK, Greene CS. Semi-supervised learning of the electronic health record for phenotype stratification. J Biomed Inform. 2016;64:168-78. https://doi.org/1/۰.۱۰۱۶j.jbi.2016.10.007

۱۷. Jothi N, Rashid NAA, Husain W. Data Mining in Healthcare – A Review. Procedia Computer Science. 2015;72:306-13. 10.1016/j.procs.2015.12.145

۱۸. Papalexakis E, Faloutsos C, Sidiropoulos ND. Tensors for Data Mining and Data Fusion: Models, Applications, and Scalable Algorithms. ACM Trans Intell Syst Technol. 2017;8(2). https://doi.org/10.1145/2915921

۱۹. Giordani P, Kiers HAL. A review of tensor-based methods and their application to hospital care data. Stat Med. 2018;37(1):137 ۵۶-https://doi.org/10.1002/sim.7514

۲۰. Ho JC, Ghosh J, Steinhubl SR, Stewart WF, Denny JC, Malin BA, et al. Limestone: high-throughput candidate phenotype generation via tensor factorization. J Biomed Inform. 2014;52:199-211. https://doi.org/10.1016/j.jbi.2014.07.001
۲۱. Luo Y, Ahmad FS, Shah SJ. Tensor Factorization for Precision Medicine in Heart Failure with Preserved Ejection Fraction. J Cardiovasc Transl Res. 2017;10(3):305-12. https://doi.org/10.1007/s12265-016-9727-8
۲۲. Zhang W, Han J, Deng S. Heart sound classification based on scaled spectrogram and tensor decomposition. Expert Syst Appl. 2017;84:220-31. https://doi.org/10.1016/j.eswa.2017.05.014
۲۳. Henderson J, He H, Malin BA, Denny JC, Kho AN, Ghosh J, et al. Phenotyping through Semi-Supervised Tensor Factorization (PSST). AMIA Annual Symposium proceedings AMIA Symposium. 2018;2018:564-73 .
۲۴. Choi J, Kim Y, Kim H-S, Choi IY, Yu H. Phenotyping of Korean patients with better-than-expected efficacy of moderate-intensity statins using tensor factorization. PLOS ONE. 2018;13(6):e0197518. https://doi.org/10.1371/journal.pone.0197518
۲۵. Sandhu R, Kaur N, Sood SK, Buyya R. TDRM: tensor-based data representation and mining for healthcare data in cloud computing environments. The Journal of Supercomputing. 2018;74(2):592-614. 10.1007/s11227-017-2163-y
۲۶. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: A systematic review. Int J Med Inform. 2018;114:57-65. https://doi.org/10.1016/j.ijmedinf.2018.03.013
۲۷. Luo J, Wu M, Gopukumar D, Zhao Y. Big Data Application in Biomedical Research and Health Care: A Literature Review. Biomed Inform Insights. 2016;8(8):1-10. 10.4137/BII.S31559
۲۸. Islam MS, Hasan MM, Wang X, Germack HD, Noor EAMA-O. A Systematic Review on Healthcare Analytics : Application and Theoretical Perspective of Data Mining. Healthcare (Basel). 2018;6(2):e54. 10.3390/healthcare6020054
۲۹. Pollard T, Johnson A, Raffa J, Celi LA, Badawi O, Mark R. eICU Collaborative Research Database (version 2.0). In: PhysioNet, editor.۲۰۱۹ .
۳۰. AL G, LAN A, L G, JM H, PCh I, RG M, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation. 2003;101(23):e215-e20 .
۳۱. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018;5(1):180178. https://doi.org/10.1038/sdata.2018.178
۳۲. Cyganek B, Woźniak M, editors. Tensor based representation and analysis of the electronic healthcare record data. IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2015 Nov 9-12 Washington, DC, USA. https://doi.org/10.1109/BIBM.2015.7359880
۳۳. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. WIREs Data Mining Knowl Discov 2020;10(4):e1379. https://doi.org/10.1002/widm.1379

۳۴. Glasser I, Sweke R, Pancotti N, Eisert J, Cirac JI. Expressive power of tensor-network factorizations for probabilistic modeling. 33rd Conference on Neural Information Processing Systems; Vancouver, Canada. p. 1496-508 .

۳۵. Cichocki A. Tensor Networks for Dimensionality Reduction, Big Data and Deep Learning. In: Gawęda AE, Kacprzyk J, Rutkowski L, Yen GG, editors. Advances in Data Analysis with Computational Intelligence Methods: Dedicated to Professor Jacek Żurada. 738. Cham: Springer International Publishing; 2018. p. 3-49.

۳۶. Scheurwegs E, Cule B, Luyckx K, Luyten L, Daelemans W. Selecting relevant features from the electronic health record for clinical code prediction. J Biomed Inform. 2017;74:92-103. https://doi.org/10.1016/j.jbi.2017.09.004

۳۷. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. Clinical Epidemiology. 2017;9:157-66. 10.2147/CLEP.S129785

۳۸. JAIN K, CHODHARY N. A SURVEY PAPER ON MISSING DATA IN DATA MINING. INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOG. 2016;3(12):45-50۳۹. Ruffini M, Gavaldà R, Limón E, editors. Clustering Patients with Tensor Decomposition. Machine Learning for Healthcare Conference 2017; Boston, MA .

۴۰. Kriegel H-P, Kröger P, Zimek A. Outlier Detection Techniques. Proceedings of the SIAM International Conference on Data Mining (SDM); Columbus .

۴۱. Skandar A, Rehman M, Anjum M. An Efficient Duplication Record Detection Algorithm for Data Cleansing. International Journal of Computer Applications. 2015;127(6):28-37. 10.5120/ijca2015906401

۴۲. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for Handling Missing Data in Electronic Health Record Derived Data. The Journal of Electronic Health Data and Methods. 2013;1(3):1035. 10.13063/2327-92۱۴.۱۰۳۵

۴۳. Endel F, Piringer H. Data Wrangling: Making data useful again. IFAC-PapersOnLine. 2015;48(1):111-2. https://doi.org/10.1016/j.ifacol.2015.05.197

۴۴. Almuhaideb S, Menai MEB. Impact of preprocessing on medical data classification. Front Comput Sci. 2016;10(6):1082-102. https://doi.org/10.1007/s11704-016-5203-5

۴۵. Hariharakrishnan J, Subramaniam M, Srividya, B. Sundhara Kumar K. Survey of pre-processing techniques for mining big data. International Conference on Computer, Communication and Signal Processing (ICCCSP); 06/08/2017; Chennai. p. 1-5. 10.1109/ICCCSP.2017.7944072