مجله دانشکده پزشکی دانشگاه علوم پزشکی مشهد
سال۶۳ ، شماره۲
خرداد-تیر ۹۹، صفحه ۲۳۶۸-۲۳۵۸

www.mums.ac.ir
Medical Journal of Mashhad
University of Medical Sciences
Vol. 63, No.2, P: 2358-2368
May-July 2020

مقاله اصلی

تحلیل داده‌های بیماران دیابتی در راستای خوشه بندی و تجویز دارو براساس الگوریتم پیشنهادی

تاریخ دریافت: ۹۸/۰۸/۰۱- تاریخ پذیرش: ۹۸/۱۲/۲٤

صفاناز حیدری[۱]
رضا رادفر[۲*]
محمود البرزی[۳]
محمدعلی افشار کاظمی[۴]
علی رجب‌زاده قطری[۵]

[۱] دانشجوی دکتری ، گروه مدیریت فناوری اطلاعات، واحد علوم و تحقیقات دانشگاه آزاد اسلامی، تهران، ایران
[۲] دانشیار ، گروه مدیریت فناوری اطلاعات، واحد علوم و تحقیقات دانشگاه آزاد اسلامی، تهران، ایران
[۳] دانشیار ، گروه مدیریت فناوری اطلاعات، واحد علوم و تحقیقات دانشگاه آزاد اسلامی، تهران، ایران
[۴] دانشیار ، گروه مدیریت صنعتی، واحد تهران مرکزی دانشگاه آزاد اسلامی، تهران، ایران
[۵] دانشیار، گروه مدیریت ، دانشکده مدیریت و اقتصاد ،دانشگاه تربیت مدرس، تهران ، ایران

Email: safanazheidari@gmail.com

خلاصه

مقدمه

دیابت یک اختلال سوخت و سازی در بدن است که توانایی تولید هورمون انسولین در بدن از بین می‌رود. هدف کلی از انجام پژوهش حاضر کشف دانش نهفته در داده‌های بیماران دیابتی است، که می‌تواند به پزشکان در خوشه‌بندی بیماران جدید و تجویز داروی مناسب مطابق هر خوشه کمک نماید.

روش کار

در این مقاله از الگوریتم MR-VDBSCAN استفاده شده است. پیاده‌سازی این الگوریتم در بستر هدوپ مبتنی بر چارچوب نگاشت-کاهش می‌باشد. ایده اصلی تحقیق استفاده از چگالی محلی برای یافتن چگالی هر نقطه است. این استراتژی می تواند مانع از اتصال خوشه‌ها با چگالی‌های متفاوت شود.

نتایج

الگوریتم موردنظر بر روی دیتاست انتخاب شده، تست و ارزیابی و نتایج نشان از دقت بالا و کارایی و مقیاس‌پذیری آن دارد. نتایج بدست آمده با نتایج اجرای خوشه‌بندی k-Means مقایسه شد، الگوریتم MR-VDBSCAN در مقایسه با آن از سرعت اجرا بالاتر ودقت تشخیص بهتری برخوردار هست و همچنین توانایی تشخیص خوشه‌ها با چگالی متفاوت برتری این الگوریتم نسبت به الگوریتم مورد مقایسه است. نتایج نشان می‌دهد که الگوریتم MR-VDBSCAN می تواند عملکرد بهتر را از سایر الگوریتم‌ها فراهم کند. به طور خاص، شباهت الگوریتم پیشنهاد شده ۹۷. برای مجموعه دیابت است.

نتیجه گیری

نتایج نشان می‌دهد که که الگوریتم MR-VDBSCAN نسبت به الگوریتم K-means خوشه-بندی بهتری را انجام می‌دهد و می‌تواند بیماران را در زیرگروههایی قرار دهد که پزشکان را در تجویز یاری نماید. نتیجه پیش‌بینی شده برای تشخیص اینکه کدوم گروه سنی و جنسیت بیشتر تحت تاثیر دیابت قرار دارند، استفاده می‌شود.

کلمات کلیدی

خوشه‌بندی، هدوپ، مپ ردیوس، داده انبوه، دیابت، داده کاوی

پی نوشت: این مطالعه فاقد تضاد منافع می‌باشد.

مقدمه

حجم انبوهی از داده‌ها در پزشکی نوین تولید ودر پایگاه داده پزشکی ذخیره می‌شود و انتظار می رود در سال‌های آینده افزایش یابد. داده‌های پزشکی اغلب ساختارنیافته هستند و شامل گزارش‌های الکترونیکی بهداشت (EHR) از داده‌های بیماران، گزارش‌های بالینی، نسخه پزشک، گزارش‌های تشخیصی، تصاویر پزشکی، اطلاعات داروسازی، داده‌های مربوط به بیمه درمانی، داده‌های مربوط به رسانه‌های اجتماعی و مجلات دارویی است(۱). ترکیب داده‌های ساختاری و بدون ساختار برای تجزیه و تحلیل پیشرفته برای بهبود نتایج پزشکی بسیار مهم است(۲). به دلیل وجوداده‌هایی که در قالب‌های نامناسب یا ناسازگار جدا شده اند یا به دلیل عدم توانایی پردازش در بارگیری و پرس و جو مجموعه داده‌های بزرگ به موقع، سازمان‌های بهداشت و درمان نمی توانند از مزایای مجموعه بزرگ داده‌های پزشکی استفاده کنند. تمامی این اطلاعات مفهوم داده انبوه را در پزشکی مطرح می‌کند. داده‌های انبوه معمولا به مجموعه‌ای از داده‌ها گفته می شود که اندازه آنها فراتر از حدی است که با ابزارهای مدیریتی و پایگاه داده معمولی بتوان آنها را در یک زمان معقول اخذ، دقیق سازی، مدیریت و پردازش کرد.به بیانی دیگر،اصطلاح داده‌های انبوه به داده‌هایی اطلاق می‌شود که به لحاظ حجم و تنوع به اندازه‌ای پیچیده هستند که مدیریت آنها با ابزارهای سنتی ممکن نبوده و از این رو نمی‌توان دا ده‌ها و دانش پنهان آنها را در زمانهای پیش‌بینی شده استخراج کرد(۳). در سال ۲۰۱۲ موسسه گارتنر تعریف جدیدی را ارایه کرد(۴): «داده‌های انبوه، حجم بالا، سرعت و شتاب بالا و تنوع بالایی از داده‌های اطلاعاتی هستند که نیازمند شکل جدیدی از پردازش هستند تا بتوانند تصمیم گیری را غنی‌تر سازند، بینش جدیدی را کشف کنند و نیز فرآیندها را بهینه نمایند.». با کمک رایانه‌های پیشرفته و بسیاری از فناوری‌های داده انبوه مانند الگوریتم‌های محاسبات ابری، هدوپ و یادگیری ماشین امکان دستیابی به کارایی بالا، مقیاس پذیری با هزینه ای نسبتاً کم امکان پذیر است. راه حل‌های بزرگ داده اغلب با مجموعه ای از راه حل‌های مدیریت داده‌های نوآورانه و ابزارهای تحلیلی همراه است، هنگامی که به طور مؤثر اجرا شوند می توانند نتایج مراقبت‌های بهداشتی را تغییر دهند(۵). داده‌های بهداشت و درمان از منابع داخلی و همچنین خارجی به سرعت در حال رشد است، از دستگاه‌های تلفن همراه، دستگاه‌های حسگر پوشیدنی، سوابق الکترونیکی بهداشت، تصاویر رادیولوژی، فیلم‌ها، یادداشت‌های بالینی، رسانه‌های اجتماعی، وبلاگ‌ها، دستگاه‌های کنترل از راه دور سلامت و غیره. سایر فرمهای پزشکی مانند تصویربرداری، خواندن سنسور نیز به مدیریت راه حل‌های داده انبوه برای مدیریت این داده‌های گسترده موجود در سازمان‌های بهداشتی کمک می کند.

داده کاوی به مفهوم استخراج اطلاعات نهان و یا الگوها و روابط مشخص در حجم زیادی از داده‌ها در یک یا چند بانک اطلاعاتی بزرگ میباشد .بسیاری از شرکتها و موسسات دارای حجم انبوهی از اطلاعات هستند. کشف دانش درون داده‌ها آن هم در عصر اطلاعات یکی از هیجان انگیزترین و کلیدی‌ترین مفاهیمی است که روز به روز اهمیت بیشتری می‌گیرد. داده-کاوی، استخراج اطلاعات مفهومی، ناشناخته و به صورت بالقوه مفید از پایگاه داده می‌باشد. خوشه‌بندی یکی از تکنیکهای با اهمیت داده‌کاوی برای کشف و استخراج دانش از دیتابیس‌های بزرگ است. خوشه‌بندی ابزار تجزیه و تحلیل اکتشافی داده‌ها با هدف گروه‌بندی اشیاء مختلف در گروه‌ها است به طوری که میزان وابستگی بین دو عنصر در یک گروه در بالاترین حد ممکن است(۶). خوشه بندی متدی است که داده‌ها در خوشه‌هایی مجزا قرار میگیرند. داده‌های درون یک خوشه بیشترین شباهت را به هم دارند و دارای ویژگی‌های مشترک هستند و دو خوشه مجزا با هم غیر مشابه هستند. خوشه-بندی با الگوریتم‌های خوشه بندی با استفاده از کامپیوتر روی دیتاستهای بزرگ امکان پذیر است و به صورت دستی انجام نمی‌گیرد.یکی از اولین برانگیزاننده‌های خوشه بندی شناسایی گروه‌های ناشناخته در درون دیتاست‌ها است(۷). صنعت پزشکی

برای بهبود نتایج خود می‌تواند با خوشه‌بندی بیماران رفتاری مناسب هریک از بیماران داشته باشند.

دیابت یا بیماری قند یک اختلال سوخت و سازی در بدن است. در این بیماری توانایی تولید هورمون انسولین در بدن از بین می‌رود یا بدن در برابر انسولین مقاوم شده و بنابراین انسولین تولیدی نمی‌تواند عملکرد طبیعی خود را انجام دهد. در افراد مبتلا به دیابت، سرعت و توانایی بدن در استفاده و سوخت و ساز کامل گلوکز کاهش می‌یابد از این‌رو میزان قند خون افزایش یافته که به آن هیپرگلیسمی می‌گویند. وقتی این افزایش قند در دراز مدت در بدن وجود داشته باشد، سبب تخریب رگ‌های بسیار ریز در بدن می‌شود که می‌تواند اعضای مختلف بدن همچون کلیه، چشم و اعصاب را درگیر کند. دیابت بیش از ۲۴۶ میلیون نفر در سراسر جهان را تحت تأثیر قرار داده که بخش اعظمی از آنها خانم هستند .طبق گزارشWHO ، تا سال ۲۰۲۵ این تعداد به بیش از ۳۸۰ میلیون نفر خواهد رسید(۸). به استثنای دیابت حاملگی – دیابت که در دوران بارداری ایجاد می شود – دو نوع اصلی وجود دارد: نوع ۱ و نوع ۲. در دیابت نوع ۱، سلولهای بتا لوزالمعده – که انسولین تولید می کنند، هورمون تنظیم کننده قند خون است – به اشتباه توسط سیستم ایمنی بدن مورد حمله و نابودی قرار می گیرند .دیابت نوع ۲ شایعترین نوع است که حدود ۷۵ تا ۸۵٪ از کل موارد را تشکیل می دهد. این اتفاق زمانی می افتد که سلول‌های بدن پاسخ خود را به انسولین متوقف می کنند، یا سلول‌های بتا قادر به تولید مقادیر کافی هورمون نیستند . این نوع خوشه‌بندی به ویژه در دیابت نوع ۲ بسیار ناهمگن است و دستوالعمل‌های پزشکی را محدود به این واقعیت است که آنها فقط به کنترل ضعیف متابولیک پاسخ می‌دهند و نمی‌توانند برای پیش‌بینی اینکه بیماران به درمان شدید نیاز دارند، مناسب باشند. خوشه‌بندی می‌تواند یک ابزار قدرتمند برای تشخیص افراد در معرض خطر و درنظر گرفتن رژیم‌های درمانی شخصی شده برای این افراد باشد و پزشکان را به سمت درمان بهینه سوق دهد. در این مقاله از الگوریتم پیشنهاد شده نویسندگان این مقاله در منبع (۹) برای خوشه‌بندی دیتاست بیماران مبتنی بر چگالی استفاده شده است. متغیرهای مورد بررسی شامل سن، جنسیت، فشارخون، شاخص توده بدن، و داشتن سابقه خانوادگی دیابت می‌باشند. پیاده‌سازی الگوریتم در چارچوب مپ ردیوس انجام گرفته و خروجی الگوریتم با الگوریتم K-means مقایسه شده است. هدف کلی از انجام پژوهش حاضر کشف دانش نهفته در داده‌های بیماران دیابتی است، که می‌تواند به پزشکان در خوشه‌بندی بیماران جدید و تجویز داروی مناسب مطابق هر خوشه کمک نماید.

روش کار

در این پژوهش تعداد ۴۵۲۳ نفر از بیماران دیابتی برای این تحلیل انتخاب شده‌اند. دیتاست انتخابی مربوط به بیمارن استان آذربایجان شرقی می‌باشد که قبلا توسط پژوهشگران دیگر بررسی و به گروه دیابت نوع ۱ و نوع ۲ تقسیم شده است که ۲۲ درصد در گروه نوع ۱ و ۷۸٪ در گروه نوع ۲ قرار گرفته‌اند. روایی دیتاست مورد بررسی توسط خبرگان این حوزه مورد تایید می‌باشد. دیتاست انتخابی دارای ۸ ویژگی است که شامل سن، جنسیت، غلظت گلوکز، فشارخون، ضخامت پوست، انسولین، شاخص توده بدن، و داشتن سابقه دیابت است. در این پژوهش با استناد به منبع (۱۰)، ۵ ویژگی (سن، جنسیت، فشارخون، شاخص توده بدن، و داشتن سابقه خانوادگی دیابت) انتخاب شدند. برای انجام تحلیل، از الگوریتم MR-VDBSCAN به عنوان الگوریتم پایه استفاده شده است (۹). با توجه به ماهیت موضوع پژوهش که در بستر مجازی انجام می‌گیرد، ابتدا نرم‌افزار VMware را نصب و دو ماشین مجازی ایجاد کرده، سپس بسته نرم‌افزاری هدوپ و جاوا و ایکلیپس را نصب می‌کنیم. روش نرم افزاری مورد استفاده در این پژوهش پردازش موازی روی نودها می‌باشد که الگوریتم خوشه‌بندی به صورت موازی اجرا شده است.

Apache Hadoop یک چارچوب نرم افزاری منبع باز مبتنی بر جاوا است که برای پردازش توزیع شده از مجموعه داده‌های بسیار زیادی در سراسر هزاران گره توزیع شده است. هدوپ داده‌ها را به قطعات کوچک تقسیم کرده و در نودها توزیع می‌کند. این نرم‌افزار متن‌باز در سال ۲۰۰۵ توسط Doug و همکارانش ارائه شده است(۱۱). آپاچی Hadoop برای

گسترش مقیاس از سرورهای تک به خوشه ای از ماشین‌های چندگانه توسعه داده شده است، هر کدام از نودها، محاسبات محلی و قابلیت ذخیره سازی خود را ارائه می دهد(۱۲). از لحاظ ساختاری هدوپ یک زیرساختار نرم افزاری برای پردازش موازی مجموعه داده‌های انبوه در کلاسترهای بزرگی از کامپیوترها است. ویژگی ذاتی هدوپ، پارتیشن‌بندی داده‌ها و پردازش موازی مجموعه داده‌های انبوه است. هدوپ بر اساس مدل برنامه‌نویسی توزیع شده Map-Reduce است که مناسب برای هر نوع داده‌ای می‌باشد. نگاشت- کاهش چارچوبی برای اجرای الگوریتم‌های توزیع‌پذیر و موازی‌پذیر در مجموعه-های داده‌ای می‌باشد(۱۳). ایده اصلی Map-Reduction این است که داده‌ها را به تکه‌هایی با اندازه ثابت تقسیم و به صورت موازی پردازش می‌کند و از مزایای آن بهره می برد. همچنین می تواند از مشکلی که هر گره کامپیوتر به طور مستقل با آن مواجه می‌شود، دوری کند(۱۴).

هدف از انجام این پژوهش خوشه‌بندی بیماران مبتلا به دیابت نوع ۲ می‌باشد. بر اساس الگوریتم MR-VDBSCAN، در لایه اول، داده‌ها از منابع مختلف جمع‌آوری و در انباره داده قرار می‌گیرند. با توجه به ماهیت پژوهش ۳۵۲۸ نفر به عنوان نمونه که دارای بیماری دیابت نوع ۲ هستند انتخاب و بر اساس ۵ ویژگی برای انجام خوشه بندی وارد مرحله دوم می‌شوند. در دومین لایه برای تقسیم‌بندی اثربخش دیتاست از الگوریتم PRBP۱ (۱۵) که تعداد نقاط حاشیه را در پارتیشن‌بندی به حداقل ممکن می‌رساند استفاده شده است. در این پارتیشن‌بندی داده‌ها به طور متوازن بین نودها توزیع شده و با کاهش تعداد نقاط مرزی کارائی خوشه‌بندی و ادغام خوشه‌های مشابه افزایش می‌یابدلایه سوم فرایند نگاشت-کاهش و ادغام و لیبل گذاری می‌باشد. تاکید این مقاله بر چگالی متنوع دیتاستهای داده‌های انبوه هست که برای این منظور از چگالی محلی هر نقطه برا جداسازی خوشه‌های با چگالی متنوع استفاده شده است و برای حل مشکل عدم پاسخگویی یک ماشین به حجم انبوه داده از نگاشت- کاهش برای پردازش داده‌ها در چندین ماشین استفاده شده است. با توجه به اینکه تعداد پارتیشن‌های ما ۲ پارتیشن است پس ما از یک مستر نود و یک اسلیو نود استفاده خواهیم کرد. در این لایه فرایند خوشه‌بندی در هر node به طور مستقل انجام می‌گیرد. هر نگاشت کننده داده‌ها را به صورت (key, value) می‌خواند که key=Null و Value=partition می‌باشند. قبل از شروع عمل خوشه‌بندی در هر نود به طور مستقل چگالی محلی هر object محاسبه و به صورت صعودی مرتب می‌شوند، نقاطی که متعلق یه یک خوشه هستند دارای مقادیر چگالی محلی نزدیک به هم خواهند بود بنابراین (فرمول۱).

فرمول ۱:

$$\text{Local-density(LDi)}=\sum_{i=1}^{K} d(x,xi)$$

نقاطی که متعلق به خوشه‌ای یکسان هستند دارای مقادیر LDi نزدیک به هم می‌باشند که می‌توان برای نقاط مجاور pi و pj در local-density list، تفاوت چگالی بین دو نقطه از طریق فرمول ۲ محاسبه می‌شود.

فرمول ۲:

$$LDVar(\text{pi,pj}) = \frac{(\text{LDp}j-\text{LDpi})}{(\text{LDpj})}$$

بعد از محاسبه تفاوت چگالی، LDVarlist را تنظیم می‌کنیم تا نقاطی که در یک سطح از خوشه قرار می‌گیرند مشخص شوند. مقادیری که در LDVar بزرگتر از حد آستانه λ باشند در یک سطح قرار می‌گیرند و در LDlevel ذخیره می‌شوند، حد آستانه با فرمول ۳ محاسبه می‌شود.

فرمول ۳: $\lambda = Ex(LDVarlist) + w.SD(LDVarlist$

EX: mathematical expectation.
SD: standard Deviation.
W: tuning coefficient (for multi-density datasets w=2.5 is a suitable value(16)

با مشخص کردن مجموعه سطوح چگالی، سطوحی که دارای چگالی مشابه باشند و تفاوت تراکم بین دو سطح کمتر از ۰/۲

[1] : Partition with Reduced boundary points

باشد (فرمول ۴) با هم ادغام می‌شوند(۱۶). در این مرحله برای هر سطح مقدار ε محاسبه(فرمول ۵) و در Eps List ذخیره می‌شوند.

فرمول ۴:

$$DenGrade\ (DLSi, DLSj) = \frac{mean\ local\ density(DLSj)}{mean\ local\ density\ (DLSi)} - 1$$

فرمول ۵:

$$\varepsilon_{i=max}\ (DLSi) = local\ density \sqrt{\frac{median\ local\ density\ (DLSi)}{mean\ local\ density\ (DLSi)}}$$

بعد از آن الگوریتم DBSCAN را با پارامتر minpts=L و مقدار EPSi از EPS list فراخوانی می‌کنیم. نتایج خوشه-بندی در دو بخش local region و boundary region تقسیم می‌شوند در فاز کاهش، الگوریتم reduce، نتایج نقاط مرزی را که از فاز mapping استخراج شده‌اند، دریافت می-کند. قبل از ارسال داده‌ها به فاز reduce در فاز shuffle ، نتایج نقاط مرزی که از فاز map بدست آمده، مورد بررسی قرار می‌گیرند و نقاط داده‌ای با qt_index یکسان از پارتیشن‌های مجاور که قابلیت ادغام در فاز reduce را دارند شناسایی و خروجی این فاز به صورت لیستی از خوشه‌هایی که قابلیت ادغام با یکدیگر را دارند، می باشد. در فاز reduce تصمیم می‌گیرد که دو خوشه‌ای که نقاط مرزی را به اشتراک گذاشته‌اند آیا با هم ادغام می‌شوند یا نه. در صورت ادغام دو خوشه، این ادغام در صورتی انجام می‌گیرد که تفاوت شعاع Eps مربوط به دو خوشه کمتر یا مساوی پارامتر θ باشد. به این صورت که از ادغام خوشه‌های با چگالی متفاوت جلوگیری می‌کنیم. در صورتی که دو خوشه دارای نقطه مرزی مشترک باشند و تفاوت شعاع Eps شان کمتر و یا مساوی θ باشد، آنگاه آن دو خوشه با هم ادغام می‌شوند. مقدار θ با توجه به کیفیت خوشه‌ای که مدنظر است می‌تواند تغییر کند. در فاز ادغام خوشه‌ها (reduce) ، خروجی فاز shuffle که لیستی از خوشه‌هایی با قابلیت ادغام است، با هم ادغام می‌شوند. خروجی این فاز لیستی از خوشه‌های ادغام شده است. در آخرین فاز اجرای الگوریتم، برچسب گذاری مجدد خوشه‌ها است. خوشه-هایی که ادغام می‌شوند برحسب کلیه cluster_id به صورت نزولی مرتب و مقدار همه خوشه‌ها با برچسب اولی لیبل‌گذاری مجدد می‌شوند. در این پژوهش تعداد خوشه‌های بدست آمده که خوشه‌هایی فرعی از دیابت نوع ۲ می‌باشند ۶ خوشه می‌باشد. برای ارزیابی دقت و کارایی از لحاظ زمان اجرا الگوریتم K-means در شرایطی مشابه برای دیتاست بیماران اجرا شد.

نتایج

پژوهش حاضر بر روی مجموعه داده ۴۵۲۳ نفری از بیماران مبتلا به دیابت انجام شده است. که قبل از عمل خوشه‌بندی با توجه به ماهیت مساله مورد بررسی در این مقاله تقریبا ۷۸٪ از دیتاست که مربوط به بیماران دارای دیابت نوع ۲ می‌باشد، انتخاب و عمل خوشه‌بندی بر مبنای ۵ فاکتور انتخاب شده انجام گرفته است. در این بررسی به کمک الگوریتم‌های MR-VDBSCAN (که توسط نویسندگان این مقاله در پژوهشی پیشین ارائه شده) و K-means، به خوشه‌بندی بیماران دیابت نوع ۲ و مقایسه نتایج آن پرداخته شده است. خلاصه‌ای از ۵ متغیر در بیماران دیابتی نوع ۲ در جدول ۱ نشان داده شده است.

جدول ۱- خلاصه ای از ۵ متغیر در بیماران دیابتی نوع ۲

بیماران دیابتی نوع ۲	بیماران دیابتی نوع ۱	فاکتورهای مورد بررسی
۳۵۲۸	۹۹۵	جنسیت
(۱۹۸۲/۱۵۴۶)		(مرد/زن)
۴۹/۵۲ ±۷/۴۶	-	سن(میانگین ±انحراف معیار)
۲۹/۲۵ ± ۳/۳۴	-	شاخص توده بدنی(میانگین ±انحراف معیار)
(۱۹۲۵/۱۶۰۳)	-	فشارخون (+/-)
(۲۸۷۵/۶۵۶)	-	داشتن سابقه خانوادگی دیابت(+/-)

با استفاده از الگوریتم MR-VDBSCAN، تعداد خوشه‌های بدست آمده ۶ خوشه‌می‌باشد که در جدول ۲ به تفکیک نشان داده شده است. همچنین از الگوریتم K-means برای خوشه-بندی مجدد داده‌ها استفاده شده، که مقایسه بین دو الگوریتم در جداول ۳ تا ۵ نشان داده شده است.

جدول۲- نتایج خوشه‌بندی توسط الگوریتم MR-VDBSCAN

	خوشه‌ها	جنسیت (مرد/زن)	سن (انحراف معیار ± میانگین)	شاخص توده بدنی (انحراف معیار ± میانگین)	فشارخون (+/-)	داشتن سابقه خانوادگی دیابت(+/-)
n=3528	Cluster1	مرد(۶۹۹)	۴۵/۳۲ ±۳/۴۶	۲۷/۳۲ ±۲/۴۶	-	-
	Cluster2	زن(۱۱۴۲)	۵۱/۵۴±۲/۸۳	۲۷/۳۵ ±۲/۶۱	-	-
	Cluster3	مرد(۳۵۷)	۴۸/۳۲ ±۷/۴۶	۲۳/۷۲ ±۲/۸۰	-	-
	Cluster4	۱۳۲/۱۷۱	۴۳/۳۲ ±۶/۲۱	۲۷/۱۶ ±۴/۳۴	-	+
	Cluster5	۴۴۲/۴۸۷	۴۴/۶۵ ±۸/۳۴	۲۶/۴۰ ±۲/۵۲	+	-
	Cluster6	۴۳/۵۵	۴۷/۵۴ ±۶/۱۲	۲۴/۳۲ ±۳/۶۳	+	+

تحلیل روی دیتاست شامل ۳۵۲۸ نمونه از بیماران که ازقبل طی بررسی‌هایی مشخص شده‌اند دارای بیماری دیابت نوع ۲ هستند، انجام شده است، خروجی کار به صورت خوشه‌هایی که بیشترین شباهت در دورن خوشه و حداقل شباهت با خوشه‌های مجاور را دارند، می‌باشد. از این تعداد ۶۹۹ نفر که مرد بودند در خوشه ۱ قرار گرفتند. خوشه ۲تعداد ۱۱۴۲ نفر از بیماران زن را شامل می‌شود. خوشه ۳ ۳۵۷ نفر مرد و خوشه‌های ۴ و ۵ و۶ شامل هر دو جنسیت می‌باشند که تعداد آنها به تفکیک در جدول ۲ آمده است. بالاترین سن مربوط به خوشه دوم و کمترین سن مربوط به خوشه چهارم می‌باشد. از لحاظ توده بدنی بیشترین وزن در خوشه‌های ۱و۲و۴ با اختلاف کمی از هم قرار دارند و کمترین وزن مربوط به خوشه سوم است. برای بررسی تکمیلی همین دیتا با الگوریتم k-means نیز اجرا شد، نتایج نشان از صحت بالاتر الگوریتم MR-VDBSCAN نسبت به K-means دارد که حاکی از عدم خوشه‌بندی صحیح برخی از مقادیر در جای مناسب خودش است.

برای مقایسه الگوریتم MR-VDBSCAN با الگوریتم K-means از ۳ معیار Rand، Jaccard و Fowlkes-Mallows که از جمله معیارهای ارزیابی خوشه‌بندی هستند. سه معیار میزان شباهت خوشه‌های حاصل از الگوریتم‌های موردارزیابی با خوشه‌های برچسب‌دار را محاسبه می‌کنند.نتایج حاصل در جداول ۳تا ۵ نشان داده شده است. برای الگوریتم‌های مورد ارزیابی با انجام تستهای مختلف بر روی مجموعه داده، بهترین پارامترها تنظیم شده است. این معیارها شباهت میان خوشه‌های به دست آمده از الگوریتم‌های ارزیابی را محاسبه می کنند نتایج ارائه شده در جداول نشان می دهد که الگوریتم MR-VDBSCAN دارای شاخص شباهت بیشتر نسبت به الگوریتم K-means است. جدول ۳، معیار جاکارد، بهترین نتایج را نسبت به K-meansنشان می دهد؛ این به این معنی است که خوشه‌های تولید شده توسط این الگوریتم بیشتر شبیه خوشه‌های برچسب‌گذاری شده هستند. شاخص Fowlkes-Mallows بر مبنای روش دوبعدی برای محاسبهTP ، TN، FPو FN است. مقدار شاخص Fowlkes-Mallows بین ۰ و ۱ است و ارزش بالا به معنای دقت بهتر است. همانطور که در جدول ۴ نشان داده شده است، ما دریافتیم که الگوریتم MR-VDBSCAN با ۹۷. برای مجموعه داده دیابت بیشترین شباهت را دارد. همچنین، برای ارزیابی اثربخشی الگوریتم پیشنهاد شده، از روش Rand برای اندازه گیری تشابه

استفاده شده است. مقدار Rand مابین [۰،۱] است و عدد نزدیک به ۱ نشان می دهد که دو پارتیشن مشابه هستند. برای ایجاد مقایسه منصفانه، ارزش یکسانی برای پارامترها در تمام الگوریتم‌ها انتخاب می کنیم، زیرا تنظیم پارامتر می تواند بر نتایج خوشه بندی و زمان اجرا تاثیر بگذارد. الگوریتم پیشنهادی شباهت زیادی دارد؛ بنابراین می توان گفت که دقت فوق العاده ای دارد و خروجی معیارها نیز تایید موضوع است، همانطور که در جدول ۵ نشان داده شده است. این جدول شباهت ۲ الگوریتم خوشه بندی را در مجموعه داده دیابت نشان می دهد. ما می بینیم که الگوریتم MR-VDBSCAN می تواند عملکرد بهتر را از سایر الگوریتم‌ها فراهم کند. به طور خاص، شباهت الگوریتم پیشنهاد شده ۹۷. برای مجموعه دیابت است.

جدول ۳- نتایج آزمون Jaccard

	دیتاست دیابت
K-means	.948
MR-VDBSCAN	.97

جدول ۴- نتایج آزمون Fowlkes_Mallows_Index

	دیتاست دیابت
K-means	.867
MR-VDBSCAN	.975

جدول۵- نتایج آزمون Rand-measure

	دیتاست دیابت
K-means	.866
MR-VDBSCAN	.894

در جدول ۶ مقایسه بین دو چارچوب انجام گرفته و از لحاظ زمان اجرا دو مدل با هم مقایسه شده‌اند. با توجه به اینکه الگوریتم MR-VDBSCAN به صورت موازی پردازش را انجام می‌دهد سریعتر از الگوریتم K-means می باشد. چراکه به صورت توزیع شده انجام گرفته است.

جدول ۶-مقایسه بین زمان دو الگوریتم مورد مقایسه

	زمان اجرا
K-means	۱۴۹۱
MR-VDBSCAN	۶۷۹

بحث و نتیجه‌گیری

دیابت یکی از شایع ترین بیماری‌های مزمن است .جمعیت مبتلا هنوز در حال رشد است. با توجه به هزینه‌های اجتماعی-اقتصادی برای مدیریت طولانی مدت بیماری دیابت، شناسایی بیماران بخش مهمی از مدیریت ودرمان دیابت به ویژه از نظر پزشکان است. در مقاله حاضر سعی شده تا با خوشه‌بندی مبتنی بر چگالی و تاکید بر چگالی متفاوت داده‌ها، به خوشه‌بندی بیماران دیابت پرداخته شود تا از این طریق بینش جدیدی برای تدوین استراتژی مدیریت دقیق‌تر دیابت ارائه شود. بررسی مطالعات مشابه که به کاربرد تکنیک‌های داده‌کاوی در زمینه بیماری دیابت می‌باشد نشان می‌دهد که روش‌های بسیاری برای تشخیص بیماری دیابت انجام گرفته در حالیکه توجه اندکی به خوشه‌بندی دیابت نوع ۲ و شناسایی زیر خوشه‌های آن شده و همچنین تحقیق در این زمینه در بستر هدوپ انجام نگرفته است. با توجه به روند رو به رشد دیتاها استفاده از هدوپ و مپ ردیوس برای پردازش داده‌های انبوه بسیار کارا می‌باشد.

کردی و همکاران در سال ۱۳۹۶ در مقاله‌ای تحت عنوان "آشکارسازی و تشخیص بیماری دیابت با استفاده از خوشه‌بندی و تکنیک‌های داده‌کاوی" برای تجزیه و تحلیل اطلاعات از نرم افزار متلب و ماشین بردار پشتیبان به منظور تشخیص بیماری استفاده کردند و برای کاهش داده‌های نویزدار از دیتاست از الگوریتم k-means استفاده کردند. نتایج حاکی از صحت بالای روش پیشنهادی توسط نویسندگان با دقت ۹۴/۲٪ در مقایسه با سایر روشهای هوشمند مورد مقایسه در مقاله داشت(۱۷).ذباح و همکاران در سال ۱۳۹۷ در مقاله‌ای تحت

عنوان" تشخیص بیماری دیابت با استفاده از شبکه عصبی مصنوعی و عصبی-فازی" با بکارگیری نرم‌افزار SPSS23، برنامه نویسی در محیط متلب و بکارگیری شبکه عصبی مصنوعی به تشخیص بیماری دیبات پرداختند نتایج تحقیق نشان می‌دهد که روش مبتنی بر عصبی فازی نسبت به سایر روش‌های مورد بررسی در مقاله دقت بالاتری دارد(۱۸). عاشوری و همکاران در سال ۱۳۹۲ در مقاله‌ای تحت عنوان " استفاده از الگوریتم‌های دسته‌بندی و خوشه‌بندی برای پیش‌بینی تعداد قرص مصرفی: موردکاوی بیماری دیابت" از تکنیک‌های داده-کاوی استفاده نمودند. در این مقاله الگوریتم‌های C5.0 و CHAID روی مجموعه داده‌های بیماران دیابتی پیاده‌سازی و درخت تصمیمی برای پیش‌بینی تعداد قرص مصرفی روزانه‌ی بیماران دیابتی اتخاذ و سپس عمل خوشه‌بندی اتخاذ گردید.نتایج مطالعه نشان می‌دهد که درخت تصمیم تولید شده برای پیش‌بینی تعداد قرص مصرفی توسط الگوریتم C5.0 از صحت بالاتری برخوردار بوده و برای پیش‌بینی مناسب‌تر است(۱۹). Durairaj و همکارانش تکنیک‌های مختلف محاسبات نرم را برای پیش بینی دیابت بررسی کردند .این تحقیق یک تکنیک موثر به عنوان شبکه عصبی مصنوعی برای پیش بینی اولیه بیماری پیشنهاد و برای بهبود دقت محاسبه آنها از الگوریتم درخت تصمیم استفاده می‌کند (۲۰).

Sadhana و همکارش در سال ۲۰۱۴ در مقاله‌ای تجزیه و تحلیل دقیق از مجموعه داده‌های دیابتی را با کمک Hiveو R. به طور کارآمد انجام دادند. واقعیت‌های موجود در طی فرآیند نشان داد می توان برای تهیه برخی از مدل‌های پیش بینی استفاده کرد. در این کار فقط تجزیه و تحلیل انجام می شود بلکه اطلاعاتی که نشان داده شده است که می تواند برای توسعه مدل‌های پیش بینی کارآمد مورد استفاده قرار گیرد(۲۱).

Biradar و همکارش در سال ۲۰۱۷ در مقاله‌ای الگوریتم‌های K-means، EM و سلسله مراتبی را از بعد عملکردی با استفاده از دیتاست بیماران دیابتی ارزیابی کرده‌اند. ارزیابی عملکردی بر اساس تعداد نمونه‌های خوشه ای و زمان اجرای زمان انجام شده برای خوشه بندی نمونه‌ها انجام شده، از نتایج مطالعه چنین استنباط می شود که الگوریتم K-means عملکردی بهتر در مقایسه با دو الگوریتم دیگر با استفاده از مجموعه داده‌های دیابت دارد(۲۲).

Ogbuabor و همکارش در سال ۲۰۱۸، به تجزیه و تحلیل تکنیک‌های خوشه بندی با استفاده از مجموعه داده‌های پزشکی، به منظور تعیین الگوریتم‌های مناسب که می تواند خوشه‌های گروه بهینه را بدست آورد، پرداخته‌اند. آنها دو الگوریتم DBSCAN و K-means برا برمبنای معیار Silhouette با هم مقایسه کردند. ابتدا الگوریتم K-means را با استفاده از تعدادهای مختلف خوشه (K) و معیارهای مسافت متفاوت، تجزیه و تحلیل کردند سپس، با استفاده از حداقل تعداد نقاط مورد نیاز برای تشکیل یک خوشه (minPts)و معیارهای مسافت متفاوت، الگوریتم DBSCANرا مورد تجزیه و تحلیل قرار دادند نتیجه آزمایش نشان می دهد که هر دو الگوریتمK- میانگین و DBSCAN دارای انسجام درون خوشه ای و جدایی بین خوشه ای هستند .بر اساس تجزیه و تحلیل، الگوریتم k-means در مقایسه با الگوریتم DBSCAN از نظر دقت خوشه بندی و زمان اجرا بهتر عمل می‌کند(۲۳).

با توجه به نتایج پژوهش، وجود روشی برای خوشه‌بندی دقیق‌تر بیماران مبتلا به دیابت برای کمک به پزشک برای افزایش صحت تشخیص و تجویز صحیح دارویی ضروری است. از جمله نقاط بهبود در این مقاله تفکیک خوشه‌های تودرتو است که الگوریتم K-means نمی‌توانست آنرا به خوبی انجام دهد.

امروزه ارائه دهندگان مراقبت‌های بهداشتی، پرداخت کنندگان، پزشکان داده‌های بزرگی را تولید می کنند که نیاز به تجزیه و تحلیل برای ارائه خدمات درمانی بهتر به بیماران دارد (۲۴). بیماران می توانند مراقبت‌های بهداشتی شخصی داشته باشند و به منظور کاهش هزینه‌ها، نقش عمده‌ای در تحلیل آن داشته باشد. در این مقاله اهمیت تحلیلی داده‌های انبوه در دیتاست بیماران دیابتی مشاهده شده است.

مدل MapReduce یک الگوی برنامه نویسی کارآمد برای پردازش چنین داده‌های انبوهی است، که مقادیر زیادی از

داده‌ها را به صورت موازی و به شیوه ای مطمئن و تحمل پذیر پردازش می کند.در این مقاله از دو الگوریتم برای خوشه‌بندی بیماران مبتلا به دیابت نوع ۲ استفاده شده است، نتایج نشان می-دهد که الگوریتم MR-VDBSCAN نسبت به الگوریتم K-means خوشه-بندی بهتری را انجام می‌دهد و می‌تواند بیماران را در زیرگروه‌هایی قرار دهد که پزشکان را در تجویز یاری نماید.

تعارض منافع

این مطالعه هیچگونه تضاد منافعی ندارد.

تشکر و قدردانی

این مطالعه به صورت مستقل و بدون حمایت مالی هیچ سازمانی انجام گرفت، پژوهشگران از کلیه افرادی که در این مطالعه همکاری نمودند بویژه از جناب آقای دکتر محمدرضا مبصری تشکر و قدردانی می‌نمایند.

حامی مالی

این مقاله برگرفته از پایاننامه دکتری صفاناز حیدری ،مدیریت فناوری اطلاعات، دانشکده مدیریت واقتصاد، دانشگاه آزاد اسلامی واحد علوم تحقیقات است. هزینه‌های این مطالعه به صورت شخصی تأمین شده است.

ملاحظات اخلاقی

در تمام مراحل کار ملاحظات اخلاقی در نظر گرفته شده است.

References

1. Kordi F, Esfandi A, Hemmati F. Detection and Diagnosis of Diabetes Using Clustering and Data Mining Techniques. The Second International Conference on Compounds, Cryptography and Computing. 2017.
2. Zabah A, Eskandi Z, Sardari A, Noghandi A. Diagnosis of diabetes using artificial and neural-fuzzy neural network. Journal of Torbat-Heydariyeh University of Medical Sciences. 2018;6(2):11-19.
3. Ashuri M, Naji Moghaddam S, Alizadeh V, Safi M. Using Classification and Clustering Algorithms to Predict the Number of Pills Taken: A Case Study of Diabetes. Health Information Management.2013;10(5):739-748.
4. Durairaj M, Kalaiselvi G. Prediction of diabetes using soft computing techniques-A survey. International journal of scientific & technology research. 2015 Mar;4(3):190-2.
5. Sharmila K, Manickam S. Diagnosing diabetic dataset using Hadoop and k-means clustering techniques. Indian Journal of Science and Technology. 2016 Oct;9(40):1-5.
6. Ahmed KN, Razak TA. An overview of various improvements of DBSCAN algorithm in clustering spatial databases. Int. J. Adv. Res. Comput. Commun. Eng.(IJARCCE). 2016 Feb;5(2):360-3.
7. Muni Kumar N, Manjula R. Role of Big data analytics in rural health care-A step towards svasth bharath. International Journal of Computer Science and Information Technologies. 2014;5(6):7172-8.
8. Ahmadi P, Sultan Aghaei M. Implementing the Hadoop Code Framework and Examining the Reduction Mapping Service. The First National Conference on New Ideas in Electrical Engineering. 2012.
9. Sohrabi B, Hamideh I. Macro Data Management in the Private and Public Sectors. Samt Publications.2015.
10. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011 Jun 9.
11. Kalyankar GD, Poojara SR, Dharwadkar NV. Predictive analysis of diabetic patient data using

machine learning and Hadoop. In2017 international conference on I-SMAC (IoT in social, mobile, analytics and cloud)(I-SMAC) 2017 Feb 10 (pp. 619-624). IEEE.

12. Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. IEEE journal of biomedical and health informatics. 2015 Feb 24;19(4):1209-15.
13. Sadhana SS, Shetty S. Analysis of diabetic data set using hive and r. International Journal of Emerging Technology and Advanced Engineering. 2014 Jul;4(7):626-9.
14. `Biradar U, Mugali DS. Clustering Algorithms on Diabetes Data: Comparative Case Study. International Journal of Advanced Research in Computer Science. 2017 May 1;8(5).
15. Ogbuabor G, Ugwoke FN. Clustering algorithm for a healthcare dataset using silhouette score value. International Journal of Computer Science & Information Technology. 2018;10(2):27-37.
16. Cho SB, Kim SC, Chung MG. Identification of novel population clusters with different susceptibilities to type 2 diabetes and their impact on the prediction of diabetes. Scientific reports. 2019 Mar 4;9(1):1-9.
17. Heidari S, Alborzi M, Radfar R, Afsharkazemi MA, Ghatari AR. Big data clustering with varied density based on MapReduce. Journal of Big Data. 2019 Dec 1;6(1):77.
18. Song J, Guo C, Wang Z, Zhang Y, Yu G, Pierson JM. HaoLap: A Hadoop based OLAP system for big data. Journal of Systems and Software. 2015 Apr 1;102:167-81.
19. Hashem IA, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU. The rise of "big data" on cloud computing: Review and open research issues. Information systems. 2015 Jan 1;47:98-115.
20. He Y, Tan H, Luo W, Feng S, Fan J. MR-DBSCAN: a scalable MapReduce-based DBSCAN algorithm for heavily skewed data. Frontiers of Computer Science. 2014 Feb 1;8(1):83-99.
21. Fu X, Wang Y, Ge Y, Chen P, Teng S. Research and application of DBSCAN algorithm based on Hadoop platform. InJoint International Conference on Pervasive Computing and the Networked World 2013 Dec 5 (pp. 73-87). Springer, Cham.
22. Lu CW, Hsieh CM, Chang CH, Yang CT. An improvement to data service in cloud computing with content sensitive transaction analysis and adaptation. In2013 IEEE 37th Annual Computer Software and Applications Conference Workshops 2013 Jul 22 (pp. 463-468). IEEE.
23. Dai BR, Lin IC. Efficient map/reduce-based dbscan algorithm with optimized data partition. In2012 IEEE Fifth international conference on cloud computing 2012 Jun 24 (pp. 59-66). IEEE.
24. Xiong Z, Chen R, Zhang Y, Zhang X. Multi-density DBSCAN algorithm based on density levels partitioning. Journal of Information and Computational Science. 2012 Oct;9(10):2739-49.

Original Article

Analysis of Diabetic Patients' Data for Clustering and Prescription Drug Based on Proposed Algorithm

Received: 23/10/2019 - Accepted: 14/03/2020

Safanaz Heidari[1]
Reza Radfar[2*]
Mahmood Alborzi[3]
Mohammad Ali Afshar Kazemi[4]
Ali Rajabzadeh Ghatari[5]

[1] Candidate P.HD., Department of Information Technology Management, Science and Research Branch, Islamic Azad University, Tehran, Iran
[2] Associate professor, Department of Industrial Management, Science and Research Branch, Islamic Azad University, Tehran, Iran
[3] Associate professor, Department of Information Technology Management, Science and Research Branch, Islamic Azad University, Tehran, Iran
[4] Associate professor, Department of Industrial Management, Central Tehran Branch, Islamic Azad University, Tehran, Iran
[5] Associate professor, Department of Management, Tarbiat Modares University, Tehran, Iran

Email: safanazheidari@gmail.com

Abstract

Introduction: Diabetes is a metabolic disorder in the body that is impaired by the ability to produce insulin hormone. The main purpose of the present study is to discover the hidden knowledge in the data of diabetic patients, which can assist clinicians in clustering new patients and prescribing appropriate medication according to each cluster.

Materials and Methods: In this paper, we use MR-VDBSCAN algorithm. The implementation of this algorithm is based on the map-reduce framework of Hadoop. The main idea of the research is to use local density to find the density of each point. This strategy can prevent clusters from joining at different densities.

Results: The algorithm is based on the selected dataset, tested and evaluated, and the results show high accuracy and efficiency. The results were compared with the results of k-Means clustering, The MR-VDBSCAN algorithm has a higher execution speed than that of the algorithm and has the ability to detect clusters with different density of superiority of this algorithm than the comparable algorithm. The results show that the MR-VDBSCAN algorithm can provide better performance than other algorithms. In particular, the similarity of the proposed algorithm is 97% for the diabetes set.

Conclusion: The results show that the MR-VDBSCAN algorithm performs better clustering than the K-means algorithm and can place patients into subgroups that assist physicians in prescribing .

Key words: Data mining, clustering, Hadoop, Map-Reduce, Big data, diabetic

Acknowledgement: There is no conflict of interest.