TUMJ مجله دانشکده پزشکی، دانشگاه علوم پزشکی تهران، اردیبهشت ۱۳۹۵، دوره ۷۴، شماره ۲، صفحه‌های ۱۰۷ تا ۱۱۲ مقاله اصیل

خوشه‌بندی بیماران مبتلا به کم‌خونی با رویکرد داده‌کاوی

خدیجه دولتشاه[۱]، رسول نورالسنا[۲*]
کامران حیدری[۳]، پریا سلیمانی[٤]
روح‌اله قاسم‌پور[٥]

۱- گروه مهندسی صنایع، مدیریت سیستم و بهره‌وری، دانشکده مهندسی صنایع، دانشگاه آزاد اسلامی، واحد تهران جنوب، تهران، ایران.
۲- گروه تولید صنعتی، دانشکده مهندسی صنایع، دانشگاه علم و صنعت ایران، تهران، ایران.
۳- گروه طب اورژانس، بیمارستان لقمان حکیم، دانشکده پزشکی، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران.
٤- گروه مهندسی صنایع، سیستم‌های اقتصادی و اجتماعی، دانشکده مهندسی صنایع، دانشگاه آزاد اسلامی، واحد تهران جنوب، تهران، ایران.
٥- گروه مدیریت خدمات بهداشتی و درمانی، دانشکده پزشکی، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران.

* نویسنده مسئول: تهران، میدان رسالت، خیابان هنگام، خیابان دانشگاه، دانشگاه علم و صنعت ایران، دانشکده مهندسی صنایع، کدپستی: ۱۳۱۱٤- ۱٦۸٤٦
تلفن: ۷۳۲۲٥۰۱۷ -۰۲۱
E-mail: rassoul@iust.ac.ir

چکیده

دریافت: ۱۳۹٤/۰٤/۰۷ ویرایش: ۱۳۹٥/۰۳/۲٤ پذیرش: ۱۳۹٥/۰٤/۰٤ آنلاین: ۱۳۹٥/۰٤/۰٥

زمینه و هدف: شایع‌ترین اختلال خونی به‌ویژه در زنان، بیماری کم‌خونی است. کشف دانش از میان حجم انبوه داده‌ها از سوابق بیماران با استفاده از داده‌کاوی می‌تواند منجر به بهبود کیفیت خدمات پزشکی شود. هدف این مطالعه خوشه‌بندی بیماران کم‌خونی با استفاده از الگوریتم‌های داده‌کاوی به‌منظور تحلیل و ارزیابی وضعیت بیماران است.

روش بررسی: در این پژوهش کاربردی، داده‌های آزمایشگاهی و بالینی بیماران کم‌خونی در جمعیت زنان مورد مطالعه قرار گرفته است. داده‌های مورد بررسی از اردیبهشت ۱۳۹۲ تا اردیبهشت ۱۳۹۳ از آزمایشگاه بیمارستان‌های امام حسین (ع) و شهدای هفتم تیر شهر تهران با ٦۹۰ رکورد و ۱٥ مشخصه‌ی آزمایشگاهی و بالینی از بیماران کم‌خونی جمع‌آوری شده‌است. برای کشف ساختارهای پنهان با استفاده از الگوریتم k-medoids بیماران خوشه‌بندی شده‌اند. برای تعیین کیفیت خوشه‌بندی از شاخص سیلوئت استفاده شده است.

یافته‌ها: مشخصه‌های Red Blood Cell (RBC)، Mean corpuscular hemoglobin (MCH)، Ferritin، GI cancer، GI infection و GI surgery بر اساس فرآیند خوشه‌بندی به‌عنوان مهم‌ترین مشخصه‌های بیماران شناسایی شده‌اند. بیماران کم‌خونی با توجه به مشخصه‌هایشان در سه خوشه توزیع شده‌اند. میانگین شاخص سیلوئت (Silhouette Coefficient) برای کیفیت خوشه‌بندی ۸۰٪ است. یعنی خوشه‌بندی دارای ساختار قوی می‌باشد.

نتیجه‌گیری: نتایج نشان داد که خوشه‌بندی با کل مشخصه‌ها نتایج مناسبی را ارایه نمی‌دهد. بنابراین هر بار با تعداد متفاوتی از مشخصه‌ها خوشه‌بندی انجام شد. نتایج خوشه‌بندی وضعیت بیماران هر خوشه را مشابه و متمایز از سایر خوشه‌ها نشان می‌دهد. خوشه اول شامل بیماران کم‌خونی فقر آهن خفیف، خوشه دوم شامل بیماران کم‌خونی فقر آهن شدید و خوشه سوم بیماران با دیگر علل کم‌خونی را دربرمی‌گیرد. تقسیم‌بندی بیماران کم‌خونی می‌تواند ابزار مفید و موثر برای تحلیل و بهبود فرآیند تصمیم‌گیری پزشکان در رابطه با درمان بیماران باشد.

کلمات کلیدی: کم‌خونی، داده‌کاوی، خوشه‌بندی، تست CBC، مشخصه‌های بالینی.

مقدمه

داده‌کاوی یک رویکرد کاربردی برای کشف الگوهای جدید و پنهان در داده‌ها می‌باشد. اطلاعات زیادی در سیستم بهداشت و درمان وجود دارد. تکنیک‌های داده‌کاوی برای انواع برنامه‌های کاربردی استفاده می‌شود.[۱] در حوزه بهداشت و درمان، داده کاوی نقش مؤثری در بهبود کیفیت خدمات ایفا می‌کند.[۲] کم‌خونی یک اختلال خونی شایع است که با کاهش غلظت هموگلوبین (HB) همراه است.[۳] کار اصلی گلبول قرمز انتقال اکسیژن از ریه‌ها به دیگر اندام‌های بدن است. بخش اصلی گلبول قرمز یعنی هموگلوبین پروتیینی است که به اکسیژن استشاقی در ریه‌ها متصل و آن را به دیگر قسمت‌های بدن انتقال می‌دهد. بنابراین عامل تعیین کننده انتقال اکسیژن سطح

هموگلوبین خون است و تعیین کننده اصلی ارزش هموگلوبین تعداد نرمال گلبول‌های قرمز در گردش خون می‌باشد. هنگامی‌که عملکرد مغز استخوان مختل و تعداد تولید روزانه گلبول‌های قرمز کم‌تر از تعداد از بین رفتن گلبول‌های قرمز پیر از گردش خون باشد، نتیجه‌ی آن کم‌خونی است.[۴] اگر اختلال کم‌خونی به سرعت به کم‌خونی شدید توسعه داده شود باعث مرگ می‌شود.[۵] تعریف سازمان بهداشت جهانی (WHO) از کم‌خونی بر مبنای میزان هموگلوبین است که بر این اساس میزان هموگلوبین کمتر از ۱۳ gr/dl برای مردان بزرگسال، کمتر از ۱۲ gr/dl برای زنان بزرگسال و برای زنان باردار بزرگسال کمتر از ۱۱ gr/dl نشان دهنده‌ی کم‌خونی است.[۶]

برای شناسایی کم‌خونی و دیگر اختلالات خونی از شمارش کامل خون Complete Blood Count (CBC) استفاده می‌شود. اغلب پزشکان در ارزیابی کم‌خونی، به سطح زیر حد نرمال هموگلوبین و هماتوکریت (HCT) اکتفا می‌کنند.[۷] شیوع کم‌خونی و پیامدهای آن در جمعیت بیماران مختلف از جمله بیماران بستری در بیمارستان، بیمارانی که جراحی داشته‌اند، بیماران قلبی، بیماران سرطانی و جمعیت زنان که کم‌خونی در بین آن ها شیوع بیشتری دارد مورد ارزیابی قرار گرفته است.[۶] گرچه بررسی‌های، شیوع بالای این بیماری در بسیاری از جمعیت‌ها را نشان می‌دهد اما داده‌ها در مورد چگونگی گسترش و موفقیت مدیریت کم‌خونی توسط پزشکان مختصر و کوتاه است و مطالعات موجود اغلب نشان‌دهنده درمان محدود کم‌خونی می‌باشد.[۶] در این پژوهش برای شناسایی مهم‌ترین مشخصه‌های بیماران کم‌خونی علاوه بر داده‌های تست شمارش کامل خون که در تحقیقات پیشین استفاده شده است، داده‌های مربوط به مشخصه‌های بالینی بیماران کم‌خونی نیز در نظر گرفته می‌شود. از پایگاه داده‌های بزرگ پزشکی می توان با استفاده از تکنیک‌های داده‌کاوی به دانش دست نیافته‌ی مفیدی رسید.[۸]

برحسب این‌که در فرآیند داده‌کاوی استنتاج چه نوع دانشی از مجموعه آموزشی مورد نظر است، از روش‌های مختلف داده‌کاوی می‌توان بهره گرفت. خوشه‌بندی یک تکنیک داده‌کاوی است که از رویکرد یادگیری غیرنظارتی برای تحلیل داده‌ها استفاده می‌کند.[۹] داده‌ها براساس اصل حداکثر کردن شباهت داخل گروه‌ها و حداقل کردن شباهت بین گروه‌ها، خوشه‌بندی می‌شوند.[۱۰] برای تشخیص کم‌خونی مشخصه‌های زیادی مورد بررسی قرار می‌گیرد، اما کدام مشخصه‌ها در تشخیص کم‌خونی و تحلیل شفاف و دقیق وضعیت بیماران کم‌خونی نقش مؤثری دارند؟ بنابراین هدف از این تحقیق خوشه‌بندی بیماران کم‌خونی با استفاده از الگوریتم k-medoids با توجه به مشخصه‌های آن‌ها است که منجر به تعیین وضعیت آن‌ها می‌شود. خوشه‌بندی در تحلیل و ارزیابی وضعیت بیماران به‌منظور بهبود فرآیند تصمیم‌گیری پزشکان در رابطه با درمان کم خونی مؤثر است.

روش بررسی

با توجه به ماهیت موضوع با یک پژوهش توصیفی، اکتشافی و کاربردی روبرو هستیم. داده‌های مورد بررسی از اردیبهشت ۱۳۹۲ تا اردیبهشت ۱۳۹۳ از آزمایشگاه بیمارستان‌های امام حسین (ع) و شهدای هفتم تیر شهر تهران جمع‌آوری شد. در این مطالعه ۶۹۰ مورد از نتایج آزمایش خون و سوابق بالینی بیماران کم‌خونی در زنان، استفاده شده است که این داده‌ها شامل ۱۵ مشخصه می‌باشند. اغلب به دلیل خطاهای عملیاتی و پیاده‌سازی سیستم‌ها، داده‌های مغشوش و ناسازگار در بین داده‌های جمع‌آوری شده وجود دارد. پردازش اولیه‌ای مورد نیاز است تا مقادیر مفقوده، انحرافات و مقادیر ثبت نشده و مسائلی از این دست را در داده‌های اولیه بیابد. این پیش پردازش جهت بهبود کیفیت داده های واقعی برای داده کاوی لازم است.

پاک‌سازی داده در واقع مرحله کنترل کیفی پیش از تحلیل داده است. در این پژوهش پیش از انجام عملیات پیش‌پردازش کلیه داده‌ها از نظر صادق بودن با مقادیر امکان‌پذیر تحت نظارت کارشناس کنترل شده‌اند و مقادیر اشتباه اصلاح و رکوردهایی که اعدادی در آن‌ها اشتباه وارد شده بودند، حذف شدند. در این مرحله تعداد کل رکوردها به ۶۸۹ مورد رسید. برای آماده‌سازی داده‌ها جهت شناسایی نقاط پرت از الگوریتم خوشه بندی k-Means استفاده شد و داده‌ها به چهار خوشه تقسیم شدند که چهار داده پرت شناسایی و با تایید کارشناس حذف شدند. در این مرحله تعداد رکوردها به ۶۸۵ رسید. برای خوشه‌بندی، داده‌ها با استفاده از روش Min-Max نرمال‌سازی شدند. نرمال‌سازی تغییر مقیاس داده‌ها به گونه‌ای است که آن‌ها را به یک فاصله کوچک و معین نگاشت می‌کند و باعث می‌شود که

داده های با مقیاس بزرگ نتـایج را بـه سـمت خـود منحـرف نکننـد. اطلاعات مشخصه‌های مورد استفاده در این پژوهش در جدول ۱ و ۲ نشان داده شده‌است. روش خوشه‌بندی در این پـژوهش الگـوریتم k-medoids می‌باشد.

در این الگوریتم به جای استفاده از مرکز یـک خوشـه بـه عنـوان مرجع، می توان از medoid ها (اشیایی که در مرکزی‌ترین محـل یـک خوشه می‌باشند) استفاده کرد. این روش بر اساس اصل حداقل‌سـازی مجموع عدم شباهت‌ها میان هر شیء و شیء مرجع عمل می‌کنـد. در ایـن پـژوهش از شـاخص اعتبـارسـنجی سـیلوئت (Silhouette Coefficient) استفاده شده است. شاخص سیلوئت یکی از معیارهـای متداول اعتبارسنجی خوشه‌بنـدی اسـت کـه دو معیـار فواصـل درون خوشه‌ای و برون خوشه‌ای را هم‌زمان در نظر می‌گیرد.[10] تفسیر مقادیر مختلـف شـاخص سـیلوئت را در جـدول ۳ مـشاهده کنیـد. بـرای آماده‌سازی داده‌ها و خوشه‌بندی بیماران جهت تحلیل وضعیت آن‌هـا از نرم‌افزار آماری R، استفاده شده‌است.

یافته‌ها

با توجه به تعداد مشخصه های بیماران مورد بررسی، ابتدا بـا کـل مشخصه‌ها خوشه‌بندی به روش k-medoids به‌ترتیب با ٤ و ۳ و ۲ =k خوشه بررسی شد. با توجه بـه تحلیـل نتـایج و بـر اسـاس شـاخص سیلوئت، خوشه‌بندی با کل مشخـصه‌هـا نتیجـه‌ی مطلـوبی بـه‌دنبـال نداشت. همچنین فرآیند خوشه بندی با علایم آزمایشگاهی و بالینی به طور جداگانه انجام شد که این فرآیند نیز نتیجـه‌ی مطلـوبی نداشـت. پس از ارزیابی نتـایج، بـا نظـر کارشناسـی و بـا اسـتفاده از میـانگین شاخص سیلوئت و بینش خاص کاربران، با مشخـصه هـای شـمارش گلبول‌های قرمز (RBC)، وزن متوسـط هموگلـوبین (MCH)، فـریتین (Ferritin)، سرطان‌های دسـتگاه گـوارش (GI cancer)، عفونـت‌هـای دستگاه گوارش (GI infection) و جراحی‌های دستگاه گـوارش (GI surgery) و k=۳ خوشه با اندازه‌هـای ۳۳۱، ۲۵۵ و ۹۹ بهتـر از سـایر نتایج تشخیص داده شد. بر اسـاس معیـار سـیلوئت، متوسـط کیفیـت خوشه‌های ۱، ۲ و ۳ به‌ترتیب دارای دقت ۷۷، ۸۳ و ۷۹٪ می‌باشند. بر این اساس متوسط کیفیت خوشه‌بندی ۸۰٪ است که بر اساس جـدول ۳ دارای ساختـار قـوی می‌باشـد. خوشه اول شامل بیمارانی است که

جدول ۱: توزیع محدوده طبیعی مشخصه‌های آزمایشگاهی مورد بررسی

مشخصه	محدوده نرمال مشخصه آزمایشگاهی (برای زنان)
شمارش گلبول‌های قرمز (X10^3/mic L)	٤/٥-۲/٤
هموگلوبین (gr/dl)	١٦-١٢
هماتوکریت (٪)	٤٨-٣٦
حجم متوسط هموگلوبین (fl)	١٠٠-٨٠
وزن متوسط هموگلوبین (pg)	٣٠-٢٧
غلظت متوسط هموگلوبین (٪)	٣٥-٣٠
پهنای گلبول قرمز (٪)	٥/١٥-٥/١٠
فریتین (ng/ml)	<١٨

جدول ۱: مشخصه‌های بالینی مورد بررسی

مشخصه	نوع	طبقات
سن	عددی	بالاتر از ١٦ سال
شاخص توده بدنی	عددی	
سرطان‌های دستگاه گوارش	طبقه‌ای	مری، معده، روده، روده بزرگ
عفونت‌های دستگاه گوارش	طبقه‌ای	کرم روده، هپاتیت B، هپاتیت C، سایر عفونت‌ها
جراحی‌های دستگاه گوارش	طبقه‌ای	مری، معده، روده، روده بزرگ
عادت‌های روزمره	طبقه‌ای	سیگار، مصرف الکل، پیکا، سایر عادت‌ها
داروهای مصرفی	طبقه‌ای	داروهای استروییدی، سایر داروها

مشخصه MCH در آن‌ها نرمال و مشخصه RBC آن‌ها پایین‌تر از حـد نرمال می‌باشد.

خوشه دوم شامل بیمارانی است که مشخصه MCH آن‌ها پایین‌تر از حد نرمال و اکثرا ویژگی RBC آن‌ها پایین‌تر از حدود نرمال است. خوشه سوم دارای ویژگی MCH بالاتر از حدود نرمال و RBC پایین‌تر از حدود نرمال می‌باشند و تعدادی از بیماران با مشخصه‌های GI cancer و GI surgery در ایـن خوشه قـرار گرفتـه‌انـد. بـا تحلیـل

جدول ۲: تفسیر مقادیر میانگین شاخص سیلوئت

میانگین شاخص سیلوئت	تفسیر
۷۱-٪۱۰۰	ساختار قوی
۵۱-٪۷۰	ساختار منطقی (مناسب)
۲۵-٪۵۰	ساختار ضعیف
<٪۲۵	هیچ ساختار قابل توجهی وجود ندارد

خوشه‌های ایجاد شده به این نتیجه رسیدیم که بیماران کم‌خونی ناشی از فقر آهن بیشتر در خوشه ۱ و ۲ توزیع شده‌اند. درصد توزیع کم‌خونی ناشی از فقر آهن درخوشه اول، دوم و سوم به‌ترتیب ٪۰۹/۲۸، ٤۱/۲۹٪ و ۰۳/۳٪ می‌باشد.

بحث

با توجه به اهمیت بهداشت و درمان در جامعه به منظور افزایش کیفیت خدمات و هم‌چنین افزایش سلامت بیماران، پژوهشگران مطالعات وسیعی را در این حوزه انجام می‌دهند. بیشتر مطالعات بر روی انواع بیماری‌ها هم‌چون بیماری‌های قلبی- عروقی، کلیوی، کبد و انواع سرطان با هدف برطرف نمودن مشکلات موجود با استفاده از روش‌های رایج داده‌کاوی هم‌چون خوشه‌بندی و دسته‌بندی و یا رویکردهای ترکیبی تمرکز داشته‌اند. در این حوزه تحقیقات در زمینه بیماری کم‌خونی کم‌تر مورد توجه قرار گرفته است. بیش‌تر مطالعات شیوع این بیماری را در جمعیت‌های مختلف مورد بررسی قرار داده‌اند و تاکنون راه‌کارهایی برای تشخیص نوع کم‌خونی در جهت درمان این بیماری ارایه نشده است.

پیش‌بینی و طبقه‌بندی کم‌خونی در جمعیت بیماران توسط Sanap و همکارانش مورد تحلیل و ارزیابی قرار گرفت. آن‌ها در تحقیق خود از تکنیک‌های داده‌کاوی درخت تصمیم C4.5 و ماشین بردار پشتیبان جهت طبقه‌بندی انواع کم‌خونی بر اساس داده‌های تست CBC استفاده کردند.[11] کم‌خونی ناشی از فقر آهن در زنان باردار با استفاده از رگرسیون لجستیک آماری توسط Casanova و همکارانش پیش‌بینی شده است. آن‌ها توانستند قانون پیش‌بینی کم‌خونی ناشی از فقرآهن را با حداکثر ویژگی‌ها توسعه دهند. پژوهش‌های آن‌ها نشان داد در زنان باردار ۲۰ هفته، وقتی سطح هموگلوبین کمتر یا مساوی ۷/۹ (HB≤۹/۷) و ۱۵≤ Red cell distribution width (RDW) داشته باشند کم‌خونی ناشی از فقر آهن را با حداکثر ویژگی‌ها می‌توان پیش‌بینی کرد که ٤۳/۷۹٪ از بیماران با دقت ۸۸٪ درست طبقه‌بندی شده‌اند.[12] با استفاده از نتایج تست CBC برای تشخیص بیماری کم‌خونی فقر آهن در زنان از روش آماری شبکه‌های عصبی مصنوعی چند لایه استفاده شده است. در این تحقیق با استفاده از این مدل در تعیین کم‌خونی فقر آهن در زنان به دقت ۱٦/۹۹٪ رسیدند.[13] یک شبکه عصبی مصنوعی (ANN) و سیستم تطبیقی فازی- عصبی (ANFIS) با توجه به نتایج تست CBC برای تشخیص کم‌خونی ناشی از فقر آهن (IDA) و همچنین پیش‌بینی سطح سرم آهن توسعه داده شده است. نتایج این بررسی نشان می‌دهد که تحلیل شبکه عصبی در تشخیص IDA نسبت به مدل‌های ANFIS و رگرسیون لجستیک بهتر است. علاوه بر این نشان می‌دهد که ANN در پیش‌بینی سطح سرم آهن دقت بالا و قابل قبولی دارد.[14]

اکثر مطالعات برای شناسایی وضعیت بیماران کم‌خونی با توجه به داده‌های تست CBC صورت گرفته است. در حالی‌که سوابق بالینی می‌تواند از علل مهم کم‌خونی باشد. در این پژوهش علاوه بر مشخصه‌های آزمایشگاهی، مشخصه‌های بالینی بیماران کم‌خونی مورد توجه قرار گرفته است. نتایج، اهمیت این مشخصه‌ها را نشان می‌دهد. بر خلاف تحقیقات پیشین در این پژوهش وضعیت بیماران کم‌خونی با رویکرد خوشه‌بندی مورد تحلیل و ارزیابی قرار گرفته‌است. این بررسی نشان می‌دهد، برای ارزیابی و تحلیل وضعیت بیماران کم‌خونی سه مشخصه‌ی آزمایشگاهی RBC، MCH، Ferritin و سه مشخصه‌ی بالینی GI cancer، GI surgery و GI infection اهمیت بیشتری دارند. با توجه به شباهت‌های درونی خوشه‌ها، نظر کارشناسی و تحلیل نتایج، اکثر بیماران مبتلا به کم‌خونی ناشی از فقر آهن در خوشه‌ی اول و دوم قرار گرفته‌اند.

با در نظر گرفتن حدود مشخصه‌های RBC و MCH در این دو گروه می‌توان گفت که بیماران در خوشه اول در معرض بیماری کم‌خونی فقر آهن خفیف و خوشه دوم کم‌خونی فقر آهن شدید هستند. خوشه سوم نیز سایر علل کم‌خونی را دارند. تقسیم‌بندی بیماران کم‌خونی با رویکرد خوشه‌بندی و براساس مؤثرترین مشخصه‌ها، می‌تواند ابزار مفید و موثر برای تحلیل و تشخیص

وضعیت بیماران و بهبود فرآیند تصمیم‌گیری پزشکان در رابطه با درمان بیماران کم‌خونی باشد. در تحقیقات آتی می‌توان از سایر مشخصه‌های آزمایشگاهی و بالینی که باعث ایجاد تمایز بیشتر بین بیماران کم‌خونی جهت شناسایی وضعیت آن‌ها می‌باشد استفاده کرد. همچنین به‌منظور تقسیم‌بندی بیماران از دیگر الگوریتم‌های خوشه‌بندی استفاده کرد.

به‌منظور تحلیل وضعیت بیماران کم‌خونی از فرآیند خوشه‌بندی استفاده شده است. خوشه‌بندی بیماران مهم‌ترین مشخصه‌های آن‌ها را تعیین کرده‌است. همچنین در این مطالعه اهمیت مشخصه های بالینی بر بیماری کم‌خونی نشان داده شده است. بیماران کم‌خونی بر اساس مهم‌ترین مشخصه‌ها به سه خوشه با وضعیت‌های متفاوت تقسیم‌بندی شده‌اند. در خوشه اول بیماران کم‌خونی فقر آهن خفیف، خوشه دوم بیماران کم خونی فقر آهن شدید و در خوشه سوم بیماران سایر علل کم‌خونی را دارند. با تحلیل وضعیت بیماران کم‌خونی جدید با توجه به مهم‌ترین مشخصه‌ها می‌توان خوشه مناسب بیمار را پیش‌بینی کرد. این کار منجر به تشخیص و درمان سریع بیمار توسط پزشکان می‌شود.

سپاسگزاری: این مقاله حاصل بخشی از پایان‌نامه تحت عنوان "کشف خوشه‌های پنهان در بیماران کم‌خونی و پیش‌بینی سطح فریتین سرم با راه‌کارهای داده‌کاوی" در مقطع کارشناسی ارشد در سال ۱۳۹٤ می‌باشد که با حمایت دانشکده مهندسی صنایع دانشگاه آزاد واحد تهران جنوب اجرا شده است.

References

1. Lakshmi KR, Kumar SP. Utilization of data mining techniques for prediction and diagnosis of major life threatening diseases survivability-review. *Int J Sci Eng Res* 2013;4(6):923-32.
2. Kaur H, Wasan SK. Empirical study on applications of data mining techniques in healthcare. *J Comput Sci* 2006;2(2):194-200.
3. O'Mara NB. Anemia in patients with chronic kidney disease. *Diab Spectrum* 2008;21(1):15-19.
4. Koury MJ. Abnormal erythropoiesis and the pathophysiology of chronic anemia. *Blood Rev* 2014;28(2):49-66.
5. D'Aquila RO, Crespo C, Mate JL, Pazos J. An inference engine based on fuzzy logic for uncertain and imprecise expert reasoning. *Fuzzy Sets Syst* 2002;129(2):187-202.
6. Shander A, Goodnough LT, Javidroozi M, Auerbach M, Carson J, Ershler WB, et al. Iron deficiency anemia-bridging the knowledge and practice gap. *Transfus Med Rev* 2014;28(3):156-66.
7. Yeh JS, Cheng CH. Using hierarchical soft computing method to discriminate microcyte anemia. *Expert Syst Appl* 2005;29(3):515-24.
8. Phillips-Wren G, Sharkey P, Dy SM. Mining lung cancer patient data to assess healthcare resource utilization. *Expert Syst Appl* 2008;35(4):1611-9.
9. Wagstaff K, Cardie C, Rogers S, Schrödl S. Constrained k-means clustering with background knowledge. In: Proceedings of the 8th International Conference on Machine; 2001.
10. Suh SC, Saffer S, Adla NK. Extraction of Meaningful Rules in a Medical Database. In: Proceedings of the 7th International Conference on Machine; 2008.
11. Sanap SA, Nagori M, Kshirsagar V. Classification of anemia using data mining techniques. In: Panigrahi BK, Suganthan PN, Das S, Satapathy SC, editors. Swarm, Evolutionary, and Memetic Computing; Berlin, Heidelberg: Springer; 2011. p. 113-21.
12. Casanova BF, Sammel MD, Macones GA. Development of a clinical prediction rule for iron deficiency anemia in pregnancy. *Am J Obstet Gynecol* 2005;193(2):460-6.
13. Yılmaz Z, Bozkurt MR. Determination of women iron deficiency anemia using neural networks. *J Med Syst* 2012;36(5):2941-5.
14. Azarkhish I, Raoufy MR, Gharibzadeh S. Artificial intelligence models for predicting iron deficiency anemia and iron serum level based on accessible laboratory data. *J Med Syst* 2012;36(3):2057-61.

TUMJ Tehran University Medical Journal, May 2016; Vol. 74, No. 2: 107-112 Original Article

Clustering of patients with anemia by data mining approach

Khadijeh Dolatshah M.Sc.[1]
Rassoul Noorossana Ph.D.[2*]
Kamran Heidari M.D.[3]
Parya Soleimani Ph.D.[4]
Roohallah Ghasempour M.Sc.[5]

1- Department of Industrial Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran.
2- Department of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran.
3- Department of Emergency Medicine, Loghman Hakim Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
4- Department of Industrial Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran.
5- Department of Health Care Management, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

[*] Corresponding author: Department of Industrial Engineering, Iran University of Science and Technology, University St., Hengam Ave., Resalat Sq., Tehran, Iran.
Post Code: 16846-13114
Tel: +98- 21- 73225017
E-mail: rassoul@iust.ac.ir

Received: 28 Jun. 2015 Revised: 13 Jun. 2016 Accepted: 24 Jun. 2016 Available online: 25 Jun. 2016

Abstract

***Background*:** Anemia disease is the most common hematological disorder which most often occurs in women. Knowledge discovery from large volumes of data associated with records of the disease can improve medical services quality by data mining The goal of this study was to determining and evaluating the status of anemia using data mining algorithms.

***Methods*:** In this applied study, laboratory and clinical data of the patients with anemia were studied in the population of women. The data have been gathered during a year in the laboratory of Imam Hossein and Shohada-ye Haft-e Tir Hospitals which contains 690 records and 15 laboratory and clinical features of anemia. To discover hidden relationships and structures using k-medoids algorithm the patients were clustered. The Silhouette index was used to determine clustering quality.

***Results*:** The features of red blood cell (RBC), mean corpuscular hemoglobin (MCH), ferritin, gastrointestinal cancer (GI cancer), gastrointestinal surgery (GI surgery) and gastrointestinal infection (GI infection) by clustering have been determined as the most important patients' features. These patients according to their features have been segmented to three clusters. First, the patients were clustered according to all features. The results showed that clustering with all features is not suitable because of weak structure of clustering. Then, each time the clustering was performed with different number of features. The silhouette index average is 80 percent that shows clustering quality. Therefore clustering is acceptable and has a strong structure.

Conclusion: The results showed that clustering with all features is not suitable because of weak structure. Then, each time the clustering was performed with different number of features. The first cluster contains mild iron deficiency anemia, the second cluster contains severe iron deficiency anemia patients and the third cluster contains patients with other anemia cause.

Keywords: anemia, CBC test, clinical features, cluster analysis, data mining.