ارائه روشی مبتنی بر الگوریتم ژنتیک جهت خوشه بندی خودکار داده های مختلط عددی و دسته ای

مسعود یقینی[۱] ، مهدی ورد[۲]

چکیده

در مسائل خوشه بندی در دنیای واقعی، اغلب با مجموعه داده هایی مواجهیم که از ترکیبی از مقادیر عددی و دسته ای تشکیل شده اند. در حالیکه اغلب روشهای خوشه بندی موجود تنها بر روی داده های عددی از کارایی مناسبی برخوردارند و قابلیت استفاده بر روی داده های مختلط را ندارند. از سوی دیگر، بیشتر روشهای سنتی و رایج، نظیر روش K-means ، تعداد خوشه ها را به عنوان ورودی از کاربر طلب می کنند. اما متاسفانه در بیشتر موارد تعداد خوشه ها برای کاربر مقداری نامعلوم است و حدس زدن مقدار آن نیز به خصوص در مورد مجموعه داده های بزرگ کاری مشکل و حتی غیرممکن است. لذا در این مقاله قصد داریم تا با بهره گیری از تعریفی دقیق تر جهت اندازه گیری فاصله میان مقادیر دسته ای، روشی را برای خوشه بندی داده های مختلط ارائه نماییم که نیازی به تعیین تعداد خوشه ها به عنوان ورودی الگوریتم نداشته و قادر است همزمان با خوشه بندی داده ها، مقدار بهینه برای تعداد خوشه ها را محاسبه نماید. همچنین در روش پیشنهادی از مزایای الگوریتم ژنتیک جهت بهینه سازی تابع هدف استفاده شده است.

کلمات کلیدی

داده کاوی، خوشه بندی، داده های مختلط، الگوریتم ژنتیک

A GA-Based Algorithm for Automatic Clustering of Mixed Numeric and Categorical Data

Masoud Yaghini, Mahdi Vard

ABSTRACT

In the real world clustering problems, it is often encountered to perform cluster analysis on data sets with mixed numeric and categorical values. However, most existing clustering algorithms are only efficient for the numeric data rather than the mixed data set. In addition, traditional methods, for example, the K-means algorithm and its variants, usually ask the user to provide the number of clusters. Unfortunately, the number of clusters in general is unknown to the user and it is difficult to guess its value especially about large datasets. We use new cost function and distance measure based on co-occurrence of values and propose a new structure to present cluster centers in the chromosomes. The genetic algorithm (GA) is used to optimize the new cost function to obtain valid clustering result. The performance of this algorithm has been studied on real world and artificial data sets. Comparisons with other clustering algorithms illustrate the effectiveness of this approach.

KEYWORDS:

---

[۱] استادیار دانشگاه علم و صنعت ایران، دانشکده راه آهن

[۲] دانشجوی کارشناسی ارشد مهندسی حمل و نقل ریلی، دانشگاه علم و صنعت ایران

Data mining; Clustering; Mixed data; Genetic algorithm

۱- مقدمه

موضوعی که در این مقاله به آن پرداخته شده است، طراحی و معرفی یک روش خوشه بندی مبتنی بر الگوریتم ژنتیک است. خوشه بندی یکی از بخشهای مهم در حیطه دانش داده کاوی است که با توجه به کاربردهای فراوانی که در حل مسایل دنیای واقعی دارد، همواره مورد توجه محققان و متخصصان داده کاوی بوده و روشهای بسیار متنوعی جهت خوشه بندی داده ها توسط آنها ارائه شده است. این روشها با توجه به رویکردی که برای گروه بندی داده ها از آن استفاده می کنند، به انواع مختلفی تقسیم می شوند که از آن میان می توان به روشهای مبتنی بر افراز داده ها، روشهای سلسله مراتبی، روشهای شبکه ای، روشهای مبتنی بر تراکم و روشهای مبتنی بر مدل اشاره کرد.

از میان روشهای بسیار زیادی که برای خوشه بندی داده ها طراحی و ارائه شده است، اغلب آنها تنها به منظور خوشه بندی نوع خاصی از داده ها (داده هایی که صرفا از نوع عددی یا صرفا از نوع دسته ای باشند) طراحی شده اند و بر روی داده های مختلط که انواع مختلف عددی و دسته ای را شامل می شوند، قابلیت کاربرد ندارند. از سوی دیگر بسیاری از مسایل دنیای واقعی و اغلب پایگاههای داده ای که روشهای خوشه بندی به منظور تحلیل آنها ایجاد شده اند، داده هایی از نوع مختلط را در خود جای داده اند. بنابراین روش خوشه بندی که قابلیت کار بر روی داده های مختلط را دارا باشد، بسیار مورد توجه و حائز اهمیت خواهد بود و چنین روشی می تواند در حل بسیاری از مسایل دنیای واقعی و تحلیل پایگاههای داده سازمانهای مختلف مورد استفاده قرار گیرد.

برای طراحی روشی که بتواند علاوه بر داده های عددی، بر روی داده های دسته ای نیز قابلیت کاربرد داشته باشد ، نخستین موضوعی که باید به آن پرداخته شود طراحی معیار سنجش فاصله میان داده های دسته ای است. در این زمینه روشهای مختلفی ارائه شده است که روش هوآنگ با توجه به سادگی و کاربرد آسان، از روشهای مورد توجه در این بخش است. اما روش هوآنگ علیرغم سادگی، از دقت خوبی برخوردار نبوده و دارای نواقصی است که سبب می شود دقت عملکرد الگوریتم خوشه بندی با معیار فاصله هوآنگ پایین آمده و جوابهایی با کیفیت پایین حاصل شود. لذا در این پایان نامه سعی نموده ایم با بهره گیری از یک معیار فاصله دقیق تر، الگوریتمی را طراحی کنیم که از دقت بالاتری نسبت به دیگر روشهای خوشه بندی داده های مختلط برخوردار باشد.

یکی دیگر از بخشهای تشکیل دهنده یک روش خوشه بندی، انتخاب رویکردی جهت جستجوی فضای جواب و بهینه سازی تابع هدف و دستیابی به جواب با مقدار تابع هدف بهتر است. یکی از تکنیکهای جستجوی فضای جواب و بهینه یابی، روشهای متاهیوریستیک هستند که از آن جمله می توان به الگوریتم ژنتیک، الگوریتم مورچه و روش جستجوی تابو اشاره نمود. گروهی از محققان نیز در طراحی روش خوشه بندی خود از این روشهای متاهیوریستیک جهت بهبود جواب خود استفاده کرده اند.

از میان تکنیکهای متا هیوریستیک، الگوریتم ژنتیک با توجه به قدرت و قابلیت بالایی که در جستجوی فضای جواب دارد، از اهمیت ویژه ای برخوردار است و لذا بسیاری از روشهای خوشه بندی نیز از این روش به عنوان ابزار مورد استفاده جهت بهینه سازی تابع هدف استفاده نموده اند. اما مرور روشهایی که در این بخش ارائه شده اند نشان می دهد که تمامی آنها تنها با هدف اجرا بر روی داده های عددی طراحی شده و هیچ یک از آنها قابلیت کاربرد بر روی داده های مختلط را دارا نیستند. وجود این خلاء ما را بر آن داشت تا با بهره گیری از قابلیت های الگوریتم ژنتیک، روش خوشه بندی ای را طراحی و ارائه نماییم که از امکان کار بر روی داده های مختلط برخوردار باشد.

یکی دیگر از جنبه های نو آورانه در روش خوشه بندی طراحی شده در این مقاله، محاسبه تعداد بهینه خوشه هاست؛ به این معنی که بر خلاف بسیاری از روشهای خوشه بندی که تعداد خوشه ها را به عنوان یک مقدار ورودی از کاربر دریافت می کنند، روش پیشنهادی ما می تواند به موازات گروه بندی داده ها و تشکیل خوشه های بهینه، مقدار بهینه برای تعداد خوشه ها را نیز محاسبه و ارائه دهد.

۲- مروری بر ادبیات موضوع

۲-۱ مروری بر روشهای خوشه بندی مبتنی بر الگوریتم ژنتیک

کریشنا و مورتی [۱] یک الگوریتم ژنتیک ترکیبی را ابداع کردند که جواب بهینه کلی را برای مساله خوشه بندی با تعداد خوشه های مشخص بدست می دهد. در این شکل ابداعی از الگوریتم ژنتیک از الگوریتم کلاسیک شیب نزولی [۱] در فرآیند خوشه بندی استفاده شده است. در الگوریتم k-means مبتنی بر الگوریتم ژنتیک (GKA)، عملگر k-means به عنوان عملگر جستجو تعریف می شود و به جای عملگر تقاطع مورد استفاده قرار می گیرد. در

روش GKA عملگر جهش خاصی متناسب با مساله خوشه بندی تعریف شده است که جهش مبتنی بر فاصله[۲] نامیده می شود. با استفاده از تئوری زنجیره مارکوف اثبات می شود که روش GKA به جواب بهینه کلی همگرا می شود. یکی از مهم ترین مشکلات در روشهای خوشه بندی از نوع افراز داده ها، یافتن افرازی از داده هاست که با داشتن تعداد مشخصی خوشه، مجموع تغییرات درون خوشه ای (TWCV)[۳] را کمینه کند. کمینه سازی مقدار TWCV در روش GKA انجام می گیرد.

الگوریتم سریع k-means مبتنی بر ژنتیک (FGKA)[۴] [۲] از الگوریتم GKA الهام گرفته شده است؛ اما در بسیاری از جنبه ها نسبت به GKA بهبود داده شده است. آزمایشات نشان می دهد که اگر چه امکان همگرایی روش k-means به یک جواب بهینه محلی وجود دارد، روشهای GKA و FGKA همواره به جواب بهینه کلی همگرا می شوند؛ اما روش FGKA بسیار سریع تر از GKA عمل می کند.

الگوریتم افزاینده k-means ژنتیک[۵] (IGKA) [۳]، توسعه ای بر الگوریتم خوشه بندی قبلی، یعنی الگوریتم سریع k-means ژنتیک (FGKA) بود. روش IGKA در مواردی که احتمال جهش مقدار کوچکی باشد، نسبت به روش FGKA از عملکرد بهتری برخوردار است. روش IGKA این ویژگی برجسته را از روش FGKA به ارث برده است که همواره به جواب بهینه کلی همگرا می باشد.

ماولیک و دیگران [۴] یک روش خوشه بندی مبتنی بر الگوریتم ژنتیک پیشنهاد کردند که در آن از قابلیت جستجوی الگوریتم ژنتیک به منظور تعیین K مرکز خوشه در فضای R^N بهره گرفته می شود. در این روش مقدار K از پیش معلوم در نظر گرفته شده است و یکی از مقادیر ورودی مساله می باشد. معیاری که برای سنجش فاصله بین نقاط از آن استفاده می شود، مجموع فاصله اقلیدسی نقاط با مرکز خوشه متناظرشان است. کروموزوم ها که به صورت رشته هایی از اعداد حقیقی هستند، مراکز خوشه ها را در خود ذخیره می کنند. هال و دیگران [۵]، روشی تحت عنوان GGA[۶] ارائه نمودند که در آن از الگوریتم ژنتیک به منظور بهینه سازی تابع هدف در روش c-means دقیق و فازی استفاده شده است. در مجموعه های داده که دارای چندین نقطه اکسترمم محلی هستند، استفاده از رویکرد الگوریتم ژنتیک از حصول جوابهای با کیفیت کمتر که دلخواه ما نیستند جلوگیری می کند.

لین و دیگران [۶]، روشی را ارائه کردند که در آن مراکز خوشه ها مستقیما از میان نقاط مجموعه داده انتخاب می شود. در این روش ابتدا یک جدول جستجو ایجاد می شود و فاصله بین هر زوج از نقاط محاسبه شده و در این جدول قرار داده می شود. در نتیجه برای محاسبه تابع برازندگی نیاز به انجام محاسبات تکراری نیست و مقادیر مورد نیاز از این جدول استخراج می شود. اتخاذ این رویکرد سبب سرعت بیشتر در محاسبه تابع برازندگی و در نتیجه سریع تر شدن کل الگوریتم می گردد. در این روش، برای نمایش جوابها در کروموزوم ها از نمایش دوتایی به جای نمایش حقیقی استفاده شده است و همچنین شاخص دیویس- بولدین[۷] برای سنجش اعتبار خوشه ها مورد استفاده قرار گرفته است.

بندیوپد و ماولیک [۷] روشی را ارائه دادند که در آن همزمان با خوشه بندی داده ها، مقدار مناسب برای تعداد خوشه ها نیز تعیین می شود. نوع جدیدی از نمایش رشته ها، که از تلفیقی از نمایش حقیقی و کاراکتر # تشکیل شده است، برای نمایش تعداد متغیر خوشه ها مورد استفاده قرار گرفته است. جهت اعتبار سنجی خوشه های حاصله نیز از شاخص Davies–Bouldin استفاده شده است.

در داده کاوی، معمولا با مجموعه های بسیار بزرگ از داده ها سر و کار داریم که از تلفیقی از مقادیر عددی و دسته ای تشکیل شده اند. اما اکثر روشهای خوشه بندی موجود تنها بر روی داده های عددی از عملکرد مناسبی برخوردارند و جهت کار بر روی داده های مختلط مناسب نیستند. جی و ژینبو [۸] روشی را جهت خوشه بندی داده های مختلط ابداع نمودند که در آن اصلاحاتی بر روی تابع هزینه متداول اعمال شده است و همچنین از ماتریس پراکندگی درون خوشه ای[۸] استفاده شده است. در این روش الگوریتم ژنتیک به منظور بهینه نمودن تابع هزینه جدید و حصول خوشه بندی معتبر به کار گرفته می شود.

لیو و دیگران [۹]، یک روش خوشه بندی ترکیبی مبتنی بر الگوریتم ژنتیک تحت عنوان روش خوشه بندی HGA ارائه نمودند. این الگوریتم با بهره گیری از لیست تابو و معیار انتظار، بین تنوع در جمعیت و سرعت همگرایی هماهنگی ایجاد می کند.

چیانگ و دیگران[۱۰]، روش k-modes را که با ایجاد تغییراتی در روش k-means، به خوشه بندی داده های دسته ای می پردازد، توسعه دادند؛ به این ترتیب که با بهره گیری از الگوریتم ژنتیک، شاخصی برای سنجش عدم تشابه با نام مقیاس فاصله ژنتیک (GDM) طراحی نمودند.

ژیانگ و فوآن [۱۱] روشی تحت عنوان KFLANN را معرفی نمودند. این روش در واقع یک شبکه عصبی کوچک است که دو نوع پارامتر را به همراه دارد، پارامتر تلورانس یا δ *و پارامتر احتیاط یا* μ *. در روش* KFLANN، الگوریتم ژنتیک به عنوان یک راه حل ممکن برای جستجوی در فضای پارامترها معرفی شده است تا به نحوی کارا و اثربخش، مقادیر مناسبی برای δ *و* μ *بدست آید.*

ژیهویی و دیگران [۱۲]، روش SPMD[۹] را ابداع کردند که در آن از ترکیب الگوریتم ژنتیک و الگوریتم جستجوی محلی دیگری تحت عنوان سربالایی[۱۰] استفاده می شود. این روش ترکیبی همگرایی الگوریتم ژنتیک را بهبود بخشیده و سرعت همگرایی را افزایش می دهد. همچنین با بکارگیری این روش ترکیبی، از ایجاد بلوغ زود رس و شرایط همگرایی نامناسب جلوگیری می گردد.

ژیانگ و دیگران [۱۳] الگوریتم k-means ژنتیک وزندار[11] (GWKMA) را که تلفیقی از الگوریتم ژنتیک و الگوریتم k-means وزندار است را پیشنهاد کردند. در روش GWKMA هر کروموزوم نمایانگر یک نحوه خوشه بندی منحصر بفرد می باشد. در این روش علاوه بر سه عملگر انتخاب، تقاطع و جهش، از عملگر دیگری تحت عنوان WKMA استفاده می شود. برتری روش GWKMA نسبت به سایر روشهای خوشه بندی مبتنی بر الگوریتم ژنتیک که از عملگر WKMA استفاده نمی کنند، نشان داده شده است.

پن و ژو [۱۴] مدلی ترکیبی برای خوشه بندی بر مبنای الگوریتم ژنتیک ارائه کردند. در این مدل که HGACLUS نامیده شده است، از مزایای روش Simulated Annealing نیز برای یافتن مراکز خوشه های بهینه یا نزدیک بهینه بهره گرفته شده است. در این مدل سعی می شود تا از طریق حداکثر کردن شباهت درون خوشه ها و نیز بیشینه نمودن تمایز بین خوشه های مختلف، بهترین خوشه بندی ممکن حاصل شود. با استفاده از استراتژی اعتبار سنجی دقیق و نیز تعداد خوشه های مشخص، ثابت شده است که روش HGACLUS نسبت به سایر روشها از دقت بیشتری برخوردار است.

کاتاری و دیگران [۱۵] ، روشی را برای خوشه بندی داده ها ارائه نمودند که در آن از الگوریتم ژنتیک بهبود یافته[12] استفاده شده و عملگرهای تقاطع و جهش به شکل کاراتری تعریف شده اند. به علاوه روش جستجوی سیمپلکس نلدر-مید (NM) و روش K-means نیز در الگوریتم ارائه شده مورد استفاده قرار گرفته است تا الگوریتم ترکیبی از مزایا و پتانسیلهای هر دو روش برخوردار باشد.

۲-۵ روشهای خوشه بندی داده های مختلط

ایجاد پایگاههای داده بسیار بزرگ که شامل مشخصه های مختلط هستند، جامعه داده کاوی را بر آن داشت تا تابع هزینه ای برای حل مسایل با داده‌های مختلط ابداع نمایند؛ چرا که الگوریتم هایی که پیش از این ایجاد شده بودند تنها بر روی داده های عددی و یا دسته ای عملکرد خوبی داشتند و قابلیت کاربرد بر روی داده هایی که دارای مشخصه هایی از هر دو نوع عددی و دسته ای بودند را نداشتند.

برای غلبه بر این مشکل، برخی از استراتژی های به کار گرفته شده به شرح زیر می باشند:

۱) مشخصه های دسته ای به مقادیر عدد صحیح تبدیل شده و سپس مقیاسهای موجود برای اندازه گیری فواصل داده های عددی برای محاسبه تشابه بین هر جفت از داده ها به کار گرفته می شود. در این روش، تخصیص مقادیر عددی صحیح به مقادیر دسته ای نظیر رنگ و ... کاری بسیار مشکل است.

۲) رویکرد دیگر به این صورت است که مقادیر مشخصه های عددی را گسسته سازی می کنند و از این طریق آنها را به صورت مقادیر دسته ای در می آورند و سپس از الگوریتم خوشه بندی داده های دسته ای استفاده می نمایند. اشکال این نوع رویکرد اینست که فرآیند گسسته سازی مقادیر عددی با از دست دادن اطلاعات توام است.

لی و بیزواز [۱۶] یک الگوریتم خوشه بندی ترکیب کننده مبتنی بر تشابهات[13] را ارائه کردند که بر اساس شاخص تشابه گودال [۱۷] عمل می کرد. این الگوریتم بر روی مشخصه های عددی و دسته ای به خوبی عمل می کند، اما اشکال آن اینست که از نظر محاسباتی هزینه زیادی دارد.

هوآنگ [۱۸] تابع هزینه ای پیشنهاد کرد که مشخصه های عددی و دسته ای را به صورت مجزا در نظر می گیرد. این تابع هزینه در الگوریتم های خوشه بندی مبتنی بر جزء بندی داده ها و بر روی مجموعه های داده مختلط قابل استفاده است. نحوه عملکرد این تابع هزینه به این شکل است که تشابه بین دو عنصر از مجموعه داده ها را به صورت مجموع دو مقدار فاصله محاسبه می نماید- یکی برای مشخصه های عددی و دیگری برای مشخصه های دسته ای. از آنجا که تابع هزینه هوآنگ قابلیت کاربرد با الگوریتم های مبتنی بر جزء بندی را دارد، هزینه های محاسباتی آن در حد مناسب و قابل قبولی است.

پس از آن، هوآنگ و دیگران روش خوشه بندی k-prototypes را برای اجرا بر روی داده های مختلط ارائه کردند. در این روش وزن هر یک از مشخصه ها به صورت خودکار بر اساس افراز کنونی داده ها محاسبه می شود.

لو و دیگران [۱۹] روشی را پیشنهاد کردند که در آن مشخصه های عددی و دسته ای به صورت جداگانه خوشه بندی می شوند و از تکنیک جمع آوری شواهد برای ترکیب نتایج خوشه بندی ها و حصول خوشه بندی نهایی استفاده می شود.

هی و دیگران [۲۰]، روش قبلی خود را که به خوشه بندی داده های دسته ای می پرداخت و الگوریتم فشرده[14] نام داشت توسعه دادند و روشی را ابداع کردند که قابلیت کاربرد بر روی داده های مختلط را دارد.

احمد و دی [۲۱] تابع هزینه ای را ارائه دادند که در آن سعی شده با بهبود و رفع نقایص تابع هوآنگ، فرآیند خوشه بندی با کیفیت بیشتری صورت گرفته و جوابهای بهتری حاصل آید. نخستین تفاوت تابع هزینه ارائه شده توسط احمد با تابع هوآنگ در اینست که تابع هوآنگ برای محاسبه فاصله بین داده های دسته ای ، بر اساس تطابق یا عدم تطابق مقدار مشخصه موردنظر در دو شیء مورد بررسی، یکی از مقادیر صفر یا یک را (صفر برای تطابق و

یک برای عدم تطابق) به عنوان مقدار فاصله تخصیص می دهد. اما احمد و دی یک تابع فاصله پیوسته را تعریف نمودند که با توجه به مقادیر سایر مشخصه ها برای دو شیء مورد بررسی، مقداری بین صفر و یک را محاسبه و آنرا به عنوان فاصله دو شیء در نظر می گیرد.

۳- معرفي روش خوشه بندي پيشنهادي

روش خوشه بندی پیشنهادی در این مقاله از نوع روشهای افراز داده ها و نیز مبتنی بر الگوریتم ژنتیک است که در آن از قابلیتهای الگوریتم ژنتیک جهت یافتن جواب بهینه استفاده می شود. همچنین این روش قادر به خوشه بندی داده های مختلط می باشد. به طور کلی مهم ترین مشخصات و امتیازات روش پیشنهادی عبارتند از:

۱) قابلیت کار بر روی داده های مختلط
۲) تعیین مقدار بهینه برای تعداد خوشه ها
۳) استفاده از روش جدیدی برای تعیین فاصله بین داده های دسته ای
۴) استفاده از شاخص Davies-Bouldin به عنوان تابع برازندگی

شکل ۳-۱ فرآیند اجرای الگوریتم پیشنهادی را نمایش می دهد. در ادامه ابتدا به تشریح روش مورد استفاده برای محاسبه فاصله بین مقادیر داده های دسته ای خواهیم پرداخت و پس از آن مراحل الگوریتم ژنتیک مورد استفاده در فرآیند خوشه بندی معرفی خواهد شد.

۳-۱ روش پیشنهادي براي محاسبه فاصله بین دو مقدار از یک متغیر دسته اي

تعریف۱- فاصله بین دو مقدار x , y از متغیر A_i نسبت به متغیر A_j و یک زیر مجموعه خاص w، به صورت زیر تعریف می شود:

$$\delta_w^i(x, y) = P_i(w|x) + P_i(\sim w|y)$$ (فرمول ۱)

تعریف ۲ - فاصله بین دو مقدار x و y از مشخصه A_i نسبت به مشخصه A_j به صورت زیر تعریف می شود:

$$\delta^{ij}(x, y) = P_i(\omega|x) + P_i(\sim \omega|y) - 1$$ (فرمول ۲)

که در رابطه فوق، ω، آن زیر مجموعه ای از مقادیر A_i (w) است که به ازای آن مقدار عبارت $P_i(\omega|x) + P_i(\sim \omega|y)$ ماکزیمم گردد.

تعریف ۳- برای یک مجموعه از داده ها که از m مشخصه، شامل مشخصه های عددی و دسته ای، تشکیل شده است، و مشخصه های عددی را در آن به صورت گسسته در آورده ایم، فاصله بین دو مقدار x , y از یک مشخصه دسته ای نسبت به یکدیگر برابر خواهد بود با:

$$\delta(x, y) = (1/m - 1) \sum_{j=1\ldots m,\, j\neq i} \delta^{ij}(x, y)$$ (فرمول ۳)

۳-۱-۱ تعیین فاصله میان دو داده

فرض کنید که D۱ و D۲ دو داده از مجموعه داده های مختلط باشند که مجموعا دارای m مشخصه می باشند. هر یک از آنها را می توانیم با $D1 = \{X_1, X_2, ..., X_m\}$ و $D2 = \{Y_1, Y_2, ..., Y_m\}$ نمایش داد که m_r مشخصه اول عددی و m_c مشخصه بعدی دسته ای هستند و $m_r + m_c = m$. فاصله بین D۱ و D۲ برابر است با:

$$Dist\ (D1, D2) = \sum_{t=1}^{m_r} (w_t(X_t - Y_t))^2 + \sum_{t=1}^{m_c} (\delta(X_t, Y_t))^2$$

(فرمول ۴)

۳-۱-۲ محاسبه مراکز خوشه ها براي داده هاي مختلط

تعریف اصلاح شده مرکز خوشه که در اینجا ارائه شده است، با نحوه تعریف مراکز خوشه ها در خوشه بندی فازی شباهت هایی دارد. اما در اینجا از این نحوه تعریف برای تعریف مراکز خوشه ها در حالت خوشه بندی با مرز مشخص[۱۵] استفاده شده است.

در روش پیشنهادی برای تعریف مراکز خوشه ها، مقدار مرکزی به ازای مشخصه های عددی همچنان با مقدار میانگین نمایش داده می شود؛ اما برای مشخصه های دسته ای از نحوه نمایش متفاوتی استفاده شده است. از آنجا که در روش پیشنهادی فاصله بین دو مقدار دسته ای بر اساس توزیع کلی آنها در سراسر مجموعه داده ها تعریف می شود، این مقدار فاصله به ازای زوجهای مختلف از مقادیر، متفاوت خواهد بود. بنابر این اگر به عنوان مثال فاصله مقدار r تا مقدار s کمتر از فاصله r تا tباشد، یعنی $\delta(r,s) < \delta(r,t)$ آنگاه انتظار می رود که در یک خوشه بندی مناسب از داده ها، تعداد رخداد های همزمان[۱۶] r و s از تعداد رخداد های همزمان r و t بیشتر باشد.

با در نظر گرفتن این مطالب، مقدار مرکزی aامین مشخصه دسته ای برای خوشه C به شکل زیر محاسبه می گردد:

$$1/N_c \left\langle (N_{1,1,,c}, N_{1,2,,c}, ..., N_{1,p1,c}), (N_{2,1,c}, N_{2,2,c}, ..., N_{2,p2,c}), ..., (N_{m,1,c}, N_{m,2,c}, ..., N_{m,pm,c}) \right\rangle$$

(فرمول ۵)

در رابطه فوق، N_c تعداد داده های موجود در خوشه C را نشان می دهد، $N_{i,k,c}$ نمایانگر تعداد داده هایی در خوشه C است که مشخصه i ام آنها دارای k امین مقدار ممکن باشد، با این فرض که مشخصه i ام دارای p_i مقدار مختلف باشد. در نتیجه مرکز خوشه توزیع نسبی هر یک از مقادیر دسته ای را در خوشه مورد نظر نشان می دهد.

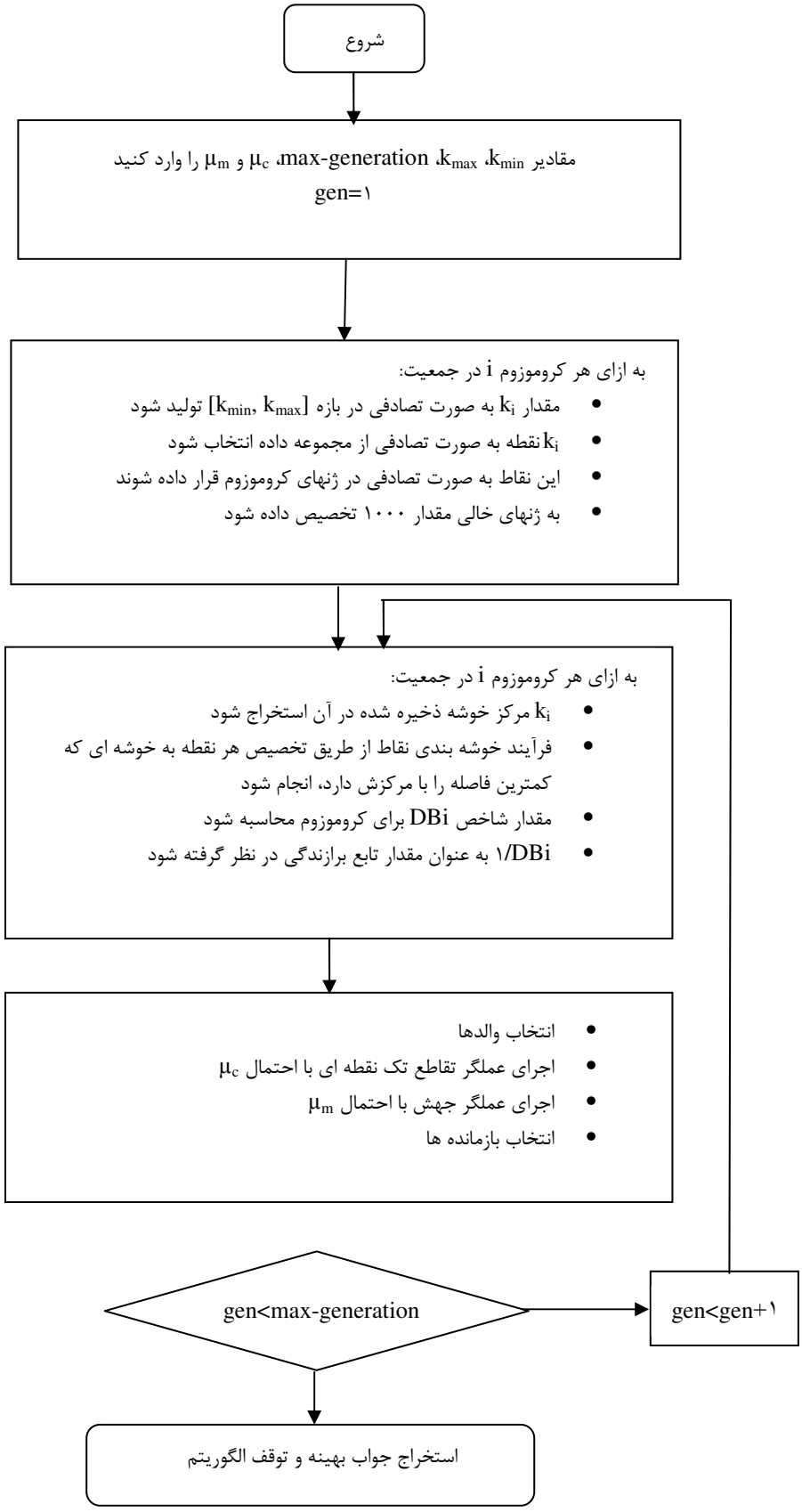
۳-۱-۳ فاصله بین یک داده و مرکز خوشه متناظرش

فاصله بین یک داده و مرکز خوشه متناظرش برابر با مجموع فواصل مقادیر عددی و دسته ای می باشد. در مورد مشخصه های عددی، فاصله اقلیدسی میان مقدار مشخصه عددی و میانگین مقادیر آن مشخصه در خوشه مورد نظر مورد استفاده قرار می گیرد. اما در مورد مشخصه های دسته ای، تمامی مقادیر ممکن آن مشخصه، همانطور که در بخش قبلی مشاهده شد، سهمی نسبی را در تعریف مرکز خوشه دارا می باشند. به ازای مشخصه دسته ای A_i ، اگر مقدار مشخصه برای داده مورد نظر برابر با r باشد، فاصله بین این داده و مرکز خوشه به صورت تابع وزنداری از مقادیر $\delta(r,v)$ محاسبه می شود که در آن، v تمامی مقادیر ممکن مشخصه A_i را اختیار می کند. از آنجا که مرکز خوشه دارای نمایشی به صورت نسبتی از تک تک مقادیر ممکن مشخصه های دسته ای می باشد، به هر یک از مقادیر فاصله $\delta(r,v)$ یک ضریب وزنی که عبارت از نسبت حضور مقدار v در خوشه است، تخصیص داده می شود.

فرض کنید $a_{i,k}$ نمایانگر kامین مقدار ممکن برای مشخصه کتگوریکال A_i باشد. همچنین فرض کنید تعداد مقادیر متمایز برای مشخصه A_i برابر با p_i باشد. با این فرضیات، فاصله به صورت رابطه زیر تعریف می گردد:

$$\Omega(X,C) = (N_{i,1,c}/N_c) * \delta(X, A_{i,1}) + (N_{i,2,c}/N_c) * \delta(X, A_{i,2}) + ... + (N_{i,p_i,c}/N_c) * \delta(X, A_{i,p_i})$$

(فرمول ۶)



(شکل ۳-۱) مراحل اجرای الگوریتم پیشنهادی

درنهایت فاصله کل میان یک داده و یک مرکز خوشه برای مجموعه داده ای شامل داده های مختلط به صورت زیر تعریف خواهد شد:

$$D\left(d_i, C_j\right) = \sum_{t=1}^{m_r} \left(d_{it}^{r} - C_{jt}^{r}\right)^2 + \sum_{t=1}^{m_c} \left(\Omega\left(d_{it}^{c}, C_{jt}^{c}\right)\right)^2$$

(فرمول ۷)

۳-۲ تبیین بخشهاي مختلف الگوریتم ژنتیک در روش پیشنهادي

۳-۲-۱ نحوه نمایش رشته ها

در روش پیشنهادی، کروموزوم ها از اعداد حقیقی تشکیل شده اند و مقادیر و مختصات مربوط به مراکز خوشه ها را در خود جای داده اند. طول رشته ها ثابت و برابر مقدار k_{max} است. مقدار k یعنی تعداد خوشه ها به صورت تصادفی از بازه $[k_{min}, k_{max}]$ انتخاب می شود که مقادیر k_{min} و k_{max} جزء ورودیهای مساله بوده که می بایست توسط کاربر معین شوند. پس از مشخص شدن مقدار k ، تعداد k ژن مراکز خوشه ها را در خود جای می دهند و به مابقی ژنها یک عدد خاص (در روش پیشنهادی ما عدد ۱۰۰۰) تخصیص داده می شود تا مشخص شود که ژن مربوطه خالی است و مرکز خوشه ای در آن قرار نگرفته است.

۳-۲-۲ مقدار دهي اولیه جمعیت

به ازای هر رشته (یا کروموزوم) iدر جمعیت ($i=۱, \ldots, P$ و P برابر با اندازه جمعیت است)، یک مقدار تصادفی k_i در بازه تعریف شده تولید می شود. سپس k_i نقطه به صورت تصادفی از میان داده ها انتخاب می شود و به صورت تصادفی در میان خانه ها رشته قرار داده می شود. در نهایت به ژنهای خالی رشته مقدار ۱۰۰۰ تخصیص داده می شود.

۳-۲-۳ تابع برازندگي

یکی دیگر از مواردی که در طراحی الگوریتم های خوشه بندی باید مدنظر قرار گیرد، انتخاب مقیاس اعتبار [۱۷] مناسب جهت انتخاب به عنوان تابع برازندگی است. شاخصهای اعتبار مختلفی نظیر شاخص Dunn، شاخص XB (شاخص Xie-Beni)، شاخص BM و شاخص DB در این زمینه ارائه شده اند.

آزمایشهای انجام شده نشان می دهد که شاخص Dunn باعث کند شدن فرآیند حل مساله می شود گرچه برای خوشه های با شکل نواری (باریکه) نتایج خوبی را به دست می دهد. شاخص XB زمانی که تعداد خوشه ها زیاد باشد از عملکرد ضعیفی برخوردار است؛ و شاخص BM در مواجهه با بیشتر مجموعه داده ها تمایل به ایجاد تنها دو خوشه دارد.

شاخص DB، که به صورت تابعی از نسبت مجموع پراکندگی نقاط در داخل خوشه به جدایی بین خوشه ها تعریف می شود، در مقایسه سایر شاخصهایی که در بالا به آنها اشاره شد نتایج معقول تری را به دست داده است.

در روش ارائه شده در این مقاله از شاخص DB به عنوان تابع برازندگی استفاده شده است. با دقت در رابطه مربوط به شاخص DB، واضح است که مقادیر کوچکتر این شاخص نشاندهنده خوشه بندی داده ها به نحوی بهتر خواهد بود. لذا از آنجا که فرآیند الگوریتم ژنتیک به دنبال بیشینه سازی تابع هدف است، معکوس این شاخص به عنوان مقدار تابع برازنگی تعریف شده است. روابط مربوطه در ادامه آورده شده است.

$$S_{i,q} = \left(\frac{1}{|C_i|} \sum_{x \in C_i} \|x - z_i\|_2^q \right)^{1/q}$$

(فرمول ۸)

$$R_{i,qt} = \underset{j, j \neq i}{Max} \left\{ \frac{S_{i,q} + S_{j,q}}{d_{ij,t}} \right\} \qquad where \quad d_{ij,t} = d(C_i, C_j) = \|z_i - z_j\|_t$$

(فرمول ۹)

$$DB_r = \frac{1}{k_r}\sum_{i=1}^{k_r} R_{i,qt}$$ (فرمول ۱۰)

$$Fitness\ (Ch_r) = \frac{1}{DB_r}$$ (فرمول ۱۱)

۴- آزمایش الگوریتم پیشنهادي

در این بخش نتایج اجرای الگوریتم خوشه بندی پیشنهادی بر روی مجموعه داده های استاندارد و داده های تصادفی تولید شده آورده شده است. و در ادامه دقت عملکرد روش پیشنهادی نسبت روشهای پیشین مقایسه گردیده است.

۴-۱ نتایج الگوریتم بر روي داده هاي استاندارد

در این قسمت نتایج اجرای الگوریتم خوشه بندی پیشنهادی بر روی مجموعه داده های استاندارد آورده شده است. به منظور مقایسه نحوه عملکرد الگوریتم طراحی شده با روشهای قبلی، از دو مجموعه داده استاندارد بهره گرفته ایم که بسیاری از روشها برای آزمایش الگوریتم خود از آن استفاده کرده اند و لذا امکان مقایسه نتایج وجود خواهد داشت. این داده ها از مخزن داده UCI [۱۸] استخراج شده اند. این داده ها به صورت از پیش طبقه بندی شده هستند و کلاس متناظر با هر رکورد مشخص شده است. در نتیجه ما برای سنجش میزان دقت الگوریتم خود، از میزان انطباق نحوه خوشه بندی داده ها با کلاسهای واقعی آنها استفاده نموده ایم. واضح است که در فرآیند خوشه بندی، ستون مربوط به اطلاعات کلاسهای داده ها در خوشه بندی لحاظ نشده و داده ها با توجه به سایر مشخصه هایشان خوشه بندی می شوند. در ادامه نتایج اجرای الگوریتم بر روی هر یک از این دو مجموعه داده استاندارد آورده شده است.

۴-۱-۱ داده هاي بيماران قلبي[۱۹]

این داده ها اطلاعات مربوط به تعدادی از بیماران قلبی را شامل می شود و در کلینیک کلولند تولید شده است. پایگاه داده اصلی شامل ۷۶ مشخصه است، اما مقالات تحقیقی مختلف برای آزمایش الگوریتم خود از یک زیر مجموعه از این پایگاه داده که شامل ۱۴ مشخصه است استفاده نموده اند. مجموعه داده های بیماران قلبی یک مجموعه داده مختلط است که از تلفیقی از داده های عددی و دسته ای تشکیل شده است؛ به این ترتیب که دارای ۹ مشخصه دسته ای و ۵ مشخصه عددی است که البته مشخصه چهاردهم کلاس داده مورد نظر رانشان می دهد و در فرآیند خوشه بندی وارد نمی شود. این مجموعه داده شامل ۳۰۳ نمونه است که در دو کلاس طبقه بندی شده اند و مجموعا ۱۶۴ نمونه متعلق به کلاس نرمال (عدم بیماری) و ۱۳۹ مورد متعلق به کلاس بیمار می باشند.

جدول ۴-۱ نتایج حاصل از خوشه بندی این مجموعه داده را به وسیله روش خوشه بندی پیشنهادی در این مقاله نشان می دهد. همچنین نتایج حاصل از پنج روش خوشه بندی داده های مختلط که برای آزمایش نتایج خود از مجموعه داده بیماران قلبی استفاده کرده اند، یعنی روشهای SBAC [۱۶] روش ECOWEB [۲۱]، روش COBWEB/۳ [۲۶] ، روش Modha & Spangler [۲۸] و روش Huang [۱۸] نشان می دهد. مقادیر دقت بدست آمده برای این الگوریتمها، دقت بیشتر و برتری روش خوشه بندی ما را نسبت به سایر روشها نشان می دهد.

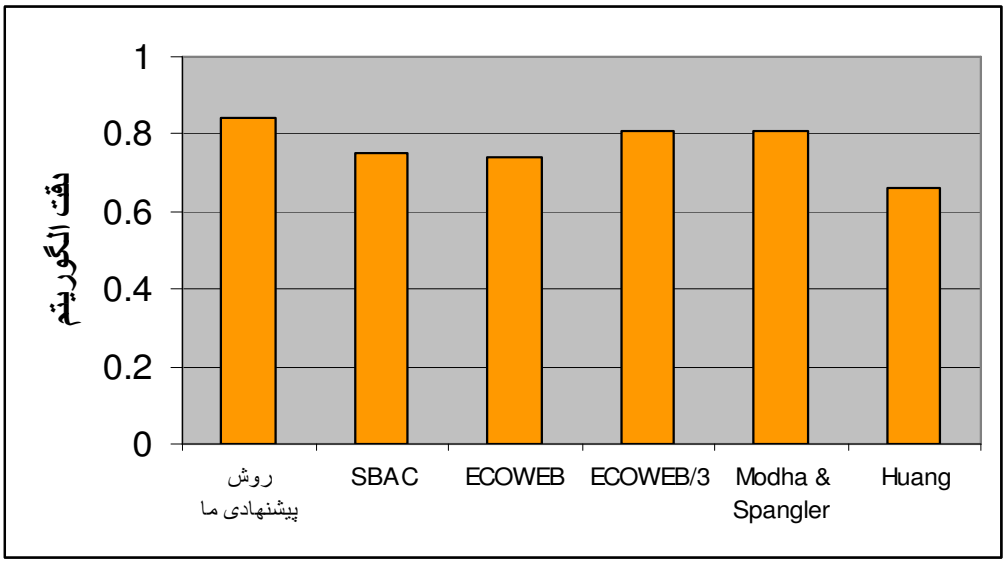
۴-۱-۲ داده هاي كارتهاي اعتباري استراليا[۲۰]

این مجموعه داده، اطلاعات مربوط به یک موسسه کارتهای اعتباری را در بردارد که در آن مشتریان به دو کلاس تقسیم بندی شده اند. این داده ها یک مجموعه داده مختلط هستند که هشت مشخصه دسته ای و شش مشخصه عددی را شامل می شوند. این مجموعه داده دارای ۶۹۰ نمونه است که به دو کلاس تقسیم می شوند: کلاس منفی شامل ۳۸۳ نمونه و کلاس مثبت شامل ۳۰۷ نمونه.

نتایج حاصل از خوشه بندی این مجموعه داده نیز توسط روش پیشنهادی ما و دو روش خوشه بندی داده های مختلط که برای آزمایش نتایج خود از مجموعه داده کارتهای اعتباری استفاده کرده اند، در جدول ۴-۲ آورده شده است. مقادیر این جدول نیز بار دیگر برتری روش پیشنهادی ما را نسبت به روش ارائه شده توسط Modha & Spangler تائید می کند.

(جدول ۴-۱) مقایسه نتایج روشهای مختلف بر روی مجموعه داده بیماران قلبی

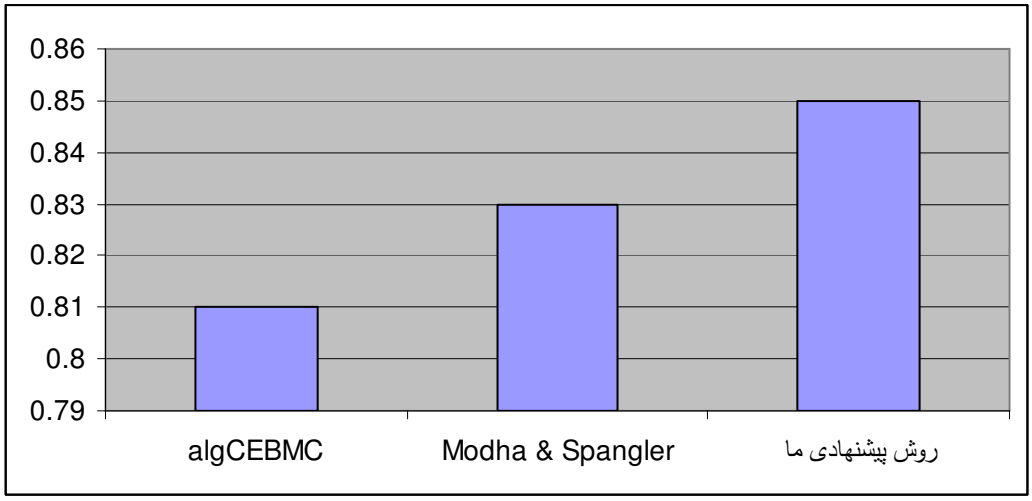
نام الگوریتم	تعداد داده هایی که در خوشه مورد انتظار قرار گرفته اند	دقت
روش پیشنهادی ما	۲۵۵	۰/۸۴
SBAC	۲۲۸	۰/۷۵
ECOWEB	۲۲۴	۰/۷۴
COBWEB/۳	۲۴۵	۰/۸۱
الگوریتم پیشنهادیModha & Spangler	۲۴۴	۰/۸۱
الگوریتم پیشنهادی Huang	۲۰۰	۰/۶۶



(نمودار ۴-۱) مقایسه میزان دقت روش خوشه بندی پیشنهادی نسبت به دیگر روشها بر روی داده های بیماران قلبی

(جدول ۴-۲) مقایسه نتایج روشهای مختلف بر روی مجموعه داده کارتهای اعتباری

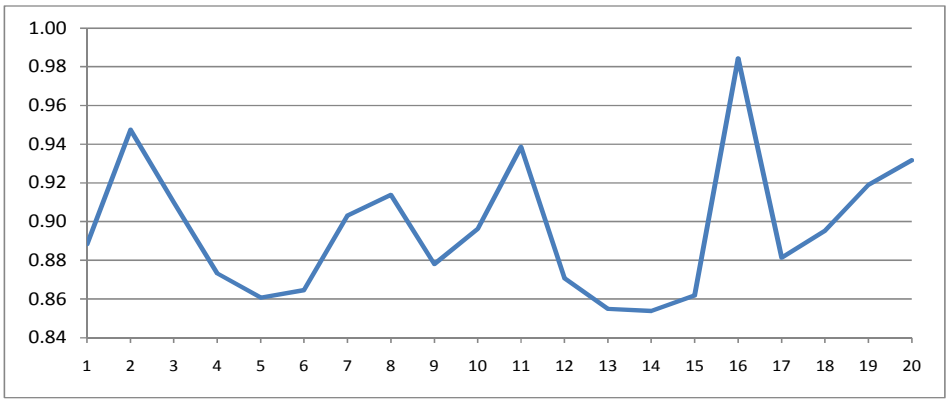
نام الگوریتم	تعداد داده هایی که در خوشه مورد انتظار قرار گرفته اند	دقت
روش پیشنهادی ما	۵۸۷	۰/۸۵
الگوریتم پیشنهادیModha & Spangler	۵۷۲	۰/۸۳
algCEBMC	۵۵۹	۰/۸۱



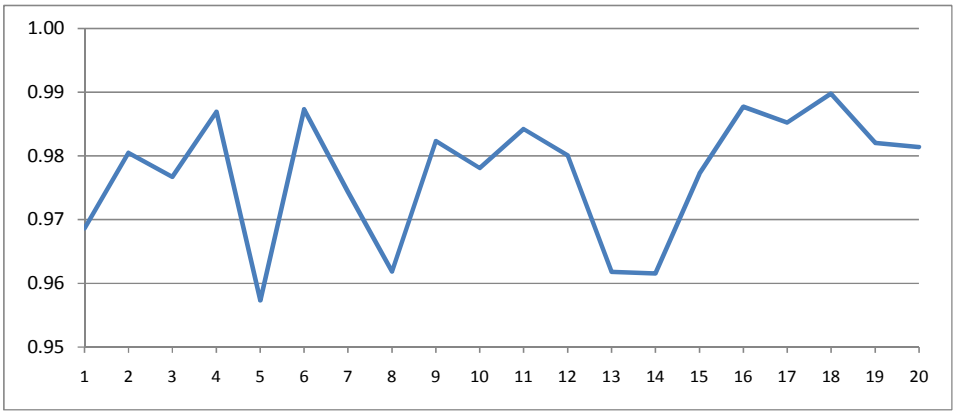
(نمودار ۴-۲) مقایسه میزان دقت روش خوشه بندی پیشنهادی نسبت به دیگر روشها بر روی داده های کارتهای اعتباری

۴-۲ نتایج الگوریتم پیشنهادي بر روي داده هاي تصادفي

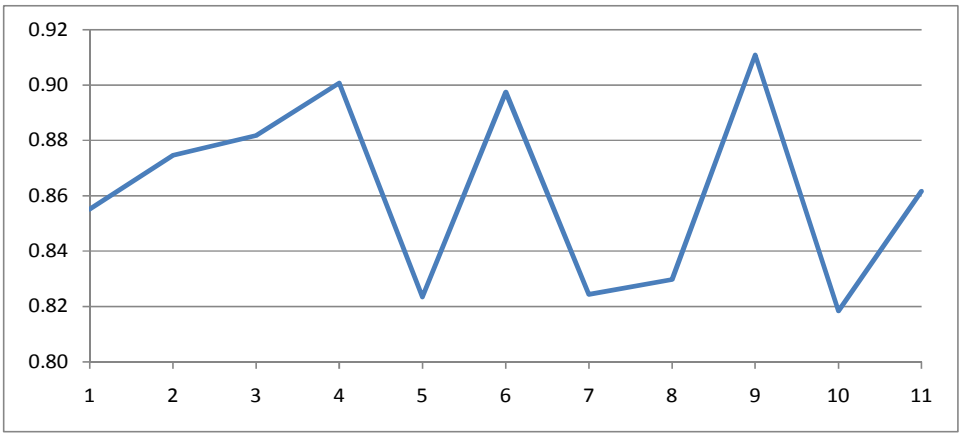
در بخش دوم جهت سنجش نتایج الگوریتم پیشنهادی، مجموعه داده هایی با ابعاد ۲۰۰۰، ۲۰۰۰۰ و ۱۲۰۰۰۰ داده تولید و فرآیند خوشه بندی در مورد آنها انجام گرفت. همچنین با تغییر نحوه تعریف فاصله بین مقادیر دسته ای در روش خوشه بندی ارائه شده توسط هوآنگ یعنی روش K-prototypes ، شکل بهبود یافته این الگوریتم نیز برای خوشه بندی این مجموعه از داده ها مورد استفاده قرار گرفته و نتایج حاصله استخراج گردید. به منظور اینکه مقایسه بین الگوریتم پیشنهادی و روش K-prototype بهبود یافته با دقت بیشتری صورت گیرد، از دو شاخص اعتبار سنجی یعنی شاخص مجموع مربعات خطاها (SSE) و شاخص دیویس-بولدین (DB) استفاده گردید و مقادیر هر دوی این شاخصها نیز برای جواب حاصل از هر یک از دو روش خوشه بندی محاسبه شد. نمودارهای ۴-۳ تا ۴-۸ نسبت مقدار شاخص DB و شاخص SSE را برای روش پیشنهادی به مقدار این دو شاخص برای روش K-prototypes نشان می دهد. نکته ای که از نمودارهای فوق الذکر قابل مشاهده است، اینست که علیرغم اینکه در روش پیشنهادی ما از شاخص DB به عنوان تابع برازندگی استفاده شده است و ما در الگوریتم خود به دنبال کمینه کردن این تابع هزینه هستیم، اما به ازای تمامی مجموعه داده های تولیدی، مقدار شاخص SSE نیز برای الگوریتم پیشنهادی مقدار کمتری را به خود اختصاص داده است.



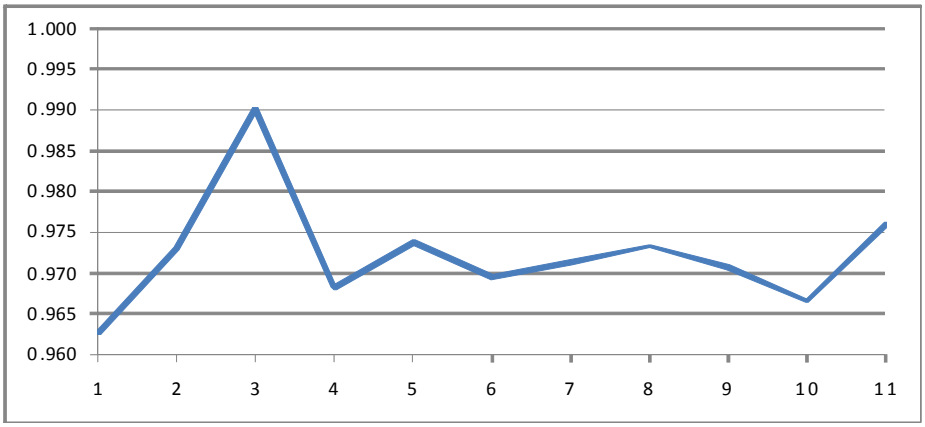
(نمودار ۴-۳) نسبت مقدار شاخص DB برای روش پیشنهادی به مقدار این شاخص برای روش K-prototypes بهبود یافته؛ مجموعه داده های ۲۰۰۰ تایی



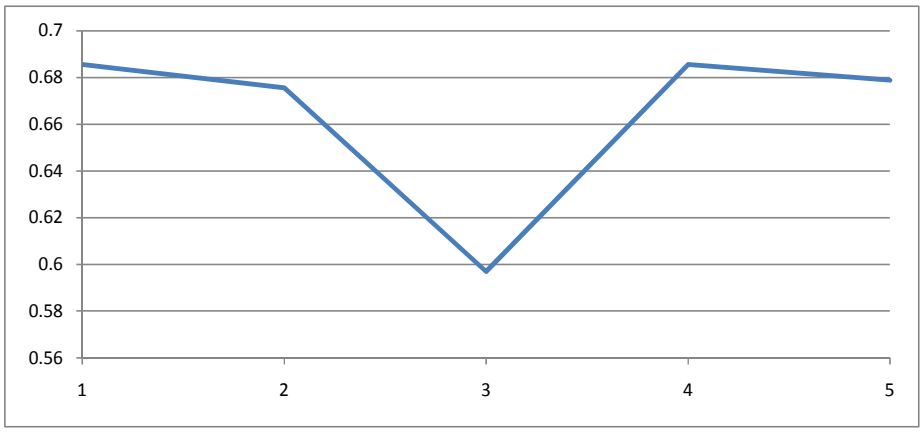
(نمودار ۴-۴) نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش K-prototypes بهبود یافته؛ مجموعه داده های ۲۰۰۰ تایی



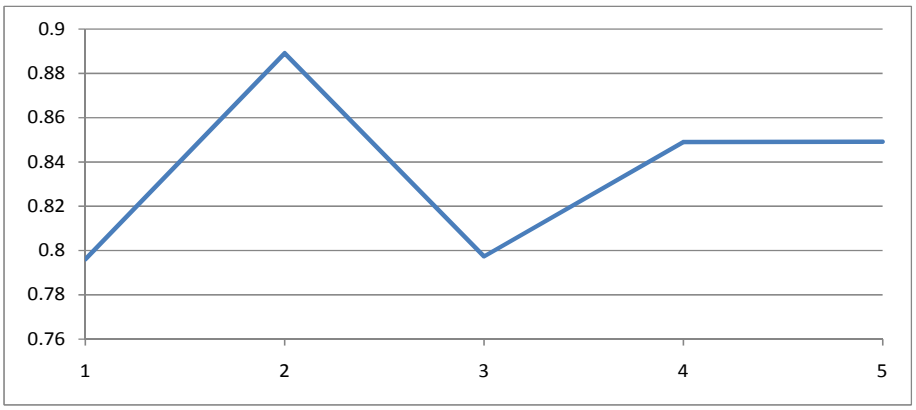
(نمودار ۴-۵) نسبت مقدار شاخص DB برای روش پیشنهادی به مقدار این شاخص برای روش K-prototypes بهبود یافته؛ مجموعه داده های ۲۰۰۰۰ تایی



(نمودار ۴-۶) نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش K-prototypes بهبود یافته؛ مجموعه داده های ۲۰۰۰۰ تایی



(نمودار ۴-۷) نسبت مقدار شاخص DB برای روش پیشنهادی به مقدار این شاخص برای روش K-prototypes بهبود یافته؛ مجموعه داده های ۱۲۰۰۰۰ تایی



(نمودار ۴-۸) نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش K-prototypes بهبود یافته؛ مجموعه داده های ۱۲۰۰۰۰ تایی

۵- نتیجه گیری

در این مقاله یک روش خوشه بندی داده های مختلط مبتنی بر الگوریتم ژنتیک ارائه شده است. ما در روش پیشنهادی خود، برخلاف اکثر روشهای خوشه بندی داده های مختلط که از معیار فاصله صفر و یک برای اندازه گیری فاصله بین داده های دسته ای بهره می برند، از تعریف جدیدی جهت سنجش فاصله بین داده های دسته ای و نیز محاسبه مراکز خوشه ها استفاده نموده ایم و سپس ساختار جدیدی را جهت نمایش این نوع از مراکز خوشه ها در خانه های کروموزوم ها ارائه کرده ایم. همچنین عملگرهای تغییر الگوریتم ژنتیک (تقاطع و جهش) متناسب با ساختار جدید نمایش مراکز خوشه ها برای هر یک از داده های عددی و دسته ای تعریف شده است. از دیگر مزایای روش پیشنهادی ما اینست که نیازی به تعیین تعداد خوش ها به عنوان ورودی الگوریتم نداشته و قادر است با بهره گیری از قابلیت جستجوی الگوریتم ژنتیک در فضای جواب، ضمن خوشه بندی داده ها، مقدار بهینه تعداد خوشه ها را نیز برای ما محاسبه نماید که این ویژگی در مورد بسیاری از مسایل دنیای واقعی که در آنها با مجموعه داده های بسیار بزرگ با تعداد خوشه های نامعین سر و کار داریم، بسیار حائز اهمیت است. آزمایش الگوریتم پیشنهادی توسط داده های استاندارد و نیز داده های مصنوعی تولید شده، نشان از برتری این روش و دقت بالاتر آن نسبت به سایر روشهای خوشه بندی داده ای مختلط دارد.

٦- مراجع

[۱]K. Krishna and M. N. Murty, "Genetic K-Means Algorithm",IEEE Transaction On Systems, Man, And cybernetics—Part B:CYBERNETICS, Vol. ۲۹, No. ۳, June ۱۹۹۹

[۲] Yi Lu, Shiyong Lu, Farshad Fotouhi ,"FGKA: A Fast Genetic K-means Clustering Algorithm", SAC'۰٤ Nicosia, Cyprus. , March ۲۰۰٤ ACM ۱-۵۸۱۱۳-۸۱۲-۱/۰۳/۰٤

[۳] Yi Lu, Shiyong Lu, Farshad Fotouhi, Youping Deng, d. Susan, J. Brown," an Incremental genetic K-means algorithm and its application in gene expression data analysis", BMCBioinformatics ۲۰۰٤

[٤]U. Maulik, S. Bandyopadhyay," Genetic algorithm-based clustering technique", Pattern Recognition ۳۳, ۲۰۰۰

[۵] Hall, L.O., Ozyurt I. B., and Bezdek, J.C, "Clustering with a genetically optimized approach". IEEE Trans. On Evolutionary Computation,۱۹۹۹

[٦] H.J. Lin, F.W. Yang and Y.T. Kao," An Efficient GAbased Clustering Technique", Tamkang Journal of Science and Engineering, Vol. ۸, No ۲, pp. ۱۱۳_۱۲۲, ۲۰۰۵

[۷]Bandyopadhyay, S. and Maulik, U., "Genetic Clustering for Automatic Evolution of Clusters and Application to Image Classification" , PatternRecognition,Vol.۳۵,pp.۱۱۹۷-۱۲۰۸, ۲۰۰۲.

[۸] LI Jie, G. Xinbo, "A GA-Based Clustering Algorithm for Large Data Sets With Mixed Numeric and Categorical Values", IEEE, Proceedings of the Fifth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'۰۳) ۰-۷٦۹۵-۱۹۵۷-۱/۰۳, ۲۰۰۳

[۹] Y. Liu, Kefe and X. Liz," A Hybrid Genetic Based Clustering Algorithm", Proceedings of the Third International Conference on Machine Leaming and Cybernetics, Shanghai, ۲٦-۲۹ August ۲۰۰٤

[۱۰] S. Chiang, S. C. Chu, Y. C. Hsin and M. H. Wang, "Genetic Distance Measure for K-Modes Algorithm", International Journal of Innovative Computing, Information and Control ICIC,ISSN ۱۳٤۹-٤۱۹۸, Volume ۲, Number ۱, pp. ۳۳-٤۰ February ۲۰۰٦

[۱۱] Yin Xiang, Alex Tay Leng Phuan," Genetic Algorithm Based K-Means Fast Learning Artificial Neural Network" , Nanyang Technological University

[۱۲] Zhihui D., Meng D., Sanli Li, Shuyou Li, Mengyue Wu and Jing Zhu, "Massively Parallel SPMD Algorithm for Cluster Computing: Combining Genetic Algorithm with Uphill.

[۱۳] Fang-Xiang Wu, Anthony J. Kusalik and W. J. Zhang, "Genetic Weighted K-means for Large-Scale Clustering Problems", University of Saskatchewan, CANADA

[۱٤] H. Pan, J. Zhu, DanfuGeno., " Genetic Algorithms Applied to Multi-Class Clustering for Gene Ex-pression Data", Geno., Prot. & Bioinfo. Vol. ۱ No. ٤ November ۲۰۰۳

[۱۵] V. Katari, S. C. Satapathy, JVR Murthy,P. Reddy ,"A Hybridized Improved Genetic Algorithm with Variable Length Chromosome for Image Clustering", International Journal of Computer Science and Network Security, VOL.۷ No.۱۱, November ۲۰۰۷

[۱٦] C. Li, G. Biswas, Unsupervised learning with mixed numeric and nominal data, IEEE Transactions on Knowledge and Data Engineering ۱٤ (٤) (۲۰۰۲) ٦۷۳–٦۹۰.

[۱۷] D.W. Goodall, A new similarity index based on probability, Biometric ۲۲ (۱۹٦٦) ۸۸۲–۹۰۷.

[۱۸] Z. Huang, Clustering large data sets with mixed numeric and Categorical values, in: Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining, World Scientific, Singapore, ۱۹۹۷.

[١٩] H. Luo, F. Kong, Y. Li, Clustering mixed data based on evidence accumulation, in: X. Li, O.R. Zaiane, Z. Li (Eds.), ADMA ٢٠٠٦, Lecture Notes on Artificial Intelligence ٤٠٩٣.

[٢٠] Z. He, X. Xu, S. Deng, Scalable algorithms for clustering large datasets with mixed type attributes, International Journal of Intelligence Systems ٢٠ (٢٠٠٥) ١٠٧٧–١٠٨٩.

[٢١] Y. Reich, S.J. Fenves, The formation and use of abstract concepts in design, in: D.H. Fisher, M.J. Pazzani, P. Langley (Eds.), Concept Formation: Knowledge and Experience in Unsupervised Learning, Morgan Kaufman, Los Altos, Calif, ١٩٩١, pp. ٣٢٣–٣٥٢.

[٢٢] J. Han and M. Kamber, Data Mining Concepts And Techniques, Elsevier, ٢٠٠٦.

[٢٣] L.H. Witten and E. Frank, Data Mining-Practical Machine Learning Tools And Techniques, Elsevier, ٢٠٠٥.

[٢٤] W.Wang, J. Yang, and R.Muntz. STING: A statistical information grid approach to spatial data mining. In Proc. ١٩٩٧ Int. Conf. Very Large Data Bases (VLDB'٩٧), pages ١٨٦–١٩٥, Athens, Greece, Aug. ١٩٩٧.

[٢٥] G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster: A multi-resolution clustering approach for very large spatial databases. In Proc. ١٩٩٨ Int. Conf. Very Large Data Bases (VLDB'٩٨), pages ٤٢٨–٤٣٩, New York, NY, Aug. ١٩٩٨.

[٢٦] K. McKusick and K. Thompson, COBWEB/٣: A Portable Implementation, Technical Report FIA-٩٠-٦-١٨-٢, NASA Ames Research Center, ١٩٩٠.

[٢٧] Zengyou, H. , Xiaofe, X. , Shengchun, D. , Clustering mixed numeric and Categorical data:A cluster ensemble approach, Department Of Computer Science And Engineering, Harbin Institute Of Technology, Harbin, China.

[٢٨] D.S. Modha, W.S. Spangler, Feature weighting in k-mean clustering, Machine Learning ٥٢ (٣) (٢٠٠٣) ٢١٧–٢٣٧.

---

[١] Gradient Descent Algorithm
[٢] Distance-based mutation
[٣] Total Within Cluster Variation
[٤] Fast Genetic K-means Algorithm
[٥] *Incremental Genetic K-means Algorithm*
[٦] Genetically Guided Algorithm
[٧] Davies-Bouldin
[٨] within cluster dispersion matrix
[٩] Single Program Multiple Data
[١٠] uphill
[١١] Genetic Weighted K-means Algorithm
[١٢] Improved Genetic Algorithm (IGA)
[١٣] Similarity Based Agglomerative Clustering (SBAC)
[١٤] Squeezed Algorithm
[١٥] Crisp Clustering
[١٦] Co-occurrence
[١٧] Validity Measure
[١٨] http://www.sgi.com/tech/mlc/db
[١٩] Heart Disease Data
[٢٠] Australian Credit Data