International Journal of Industrial Engineering & Production Management (2012)

September 2012, Volume 23, Number 2
pp. 187-197

http://IJIEPM.iust.ac.ir/

Automatic Clustering of Mixed Data Using Genetic Algorithm

M.Yaghini* **& M.Vard**

Masoud Yaghini, *Assistance professor of School of Railway Engineering - Iran University of Science and Technology*
Mahdi Vard, *MSc, School of Railway Engineering - Iran University of Science and Technology*

Keywords

Data mining
Clustering
Mixed data
Genetic algorithm
Davies-Bouldin index

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, usually ask the user to provide the number of clusters. In this paper, we propose a new method to cluster mixed data and automatically evolve the number of clusters as well as clustering of data set. In the proposed method, Davies-Bouldin Index is used as fitness function and we use the genetic algorithm to optimize fitness function. Also, we use a more accurate distance measure for calculating the distance between categorical values. The performance of this algorithm has been studied on real world and simulated data sets. Comparisons with other clustering algorithms illustrate the effectiveness of this approach.

* **Corresponding author. Masoud Yaghini**
Email: *yaghini@iust.ac.ir*

نشـــریه بین المللـــی مهنـــدسی صنـــایع و مدیـــــریت تولیـــد

شمـــــاره ۲، جلـــد ۲۳، شهریـــــور ۱۳۹۱
صفحـــــه ۱۹۷-۱۸۸

نشریه بین المللی مهندسی صنایع و مدیریت تولید
دانشگاه علم و صنعت ایران

ISSN: 2008-4870

http://IJIEPM.iust.ac.ir/

خوشه‌بندی خودکار داده‌های مختلط با استفاده از الگوریتم ژنتیک

مسعود یقینی* و مهدی ورد

کلمات کلیدی

داده‌کاوی
خوشه‌بندی
داده‌های مختلط
الگوریتم ژنتیک
شاخص Davies-Bouldin

چکیده:

مساله خوشه بندی به منظور کمینه کردن مجموع مجذور انحراف، یک مساله غیر خطی و غیر محدب بوده و دارای تعداد زیادی نقاط بهینه محلی است. در مسائل خوشه‌بندی در دنیای واقعی، اغلب با مجموعه داده‌هایی مواجهیم که از ترکیبی از مقادیر عددی و دسته‌ای تشکیل شده‌اند. در حالیکه اغلب روشهای خوشه‌بندی موجود تنها بر روی داده‌های عددی از کارایی مناسبی برخوردارند و قابلیت استفاده بر روی داده‌های مختلط را ندارند. از سوی دیگر، بیشتر روشهای سنتی، تعداد خوشه‌ها را به عنوان ورودی از کاربر طلب می‌کنند. در حالیکه در بیشتر موارد تعداد خوشه‌ها برای کاربر مقداری نامعلوم است و حدس زدن مقدار آن نیز به خصوص در مورد مجموعه داده‌های بزرگ کاری مشکل و حتی غیرممکن است. در این مقاله قصد داریم تا با بهره‌گیری از روشی دقیق‌تر جهت اندازه‌گیری فاصله میان مقادیر دسته‌ای، روش جدیدی را برای خوشه‌بندی داده‌های مختلط ارائه نماییم که نیازی به تعیین تعداد خوشه‌ها به عنوان ورودی الگوریتم نداشته و قادر است همزمان با خوشه‌بندی داده ها، مقدار بهینه برای تعداد خوشه‌ها را محاسبه نماید. در روش پیشنهادی معکوس شاخص Davies-Bouldin به عنوان تابع برازش در نظر گرفته شده و به منظور جستجوی فضای جواب از الگوریتم ژنتیک استفاده می‌شود. برای ارزیابی عملکرد الگوریتم از دو گروه از داده های استاندارد و شبیه‌سازی شده استفاده شده است. نتایج بدست آمده، عملکرد بسیار بالای الگوریتم پیشنهادی را نشان می‌دهد.

۱. مقدمه

خوشه‌بندی عبارت از دسته‌بندی تعدادی از اشیاء به منظور ایجاد گروههایی از اشیاء تحت عنوان خوشه است. به نحویکه اشیاء درون یک خوشه بسیار شبیه هم بوده و اشیاء موجود در خوشه‌های مختلف، کاملا متمایز از یکدیگر باشند. روشهای مختلف خوشه‌بندی با توجه به رویکردی که برای گروه‌بندی داده‌ها از آن استفاده می‌کنند، به انواع مختلفی تقسیم می‌شوند که از آن میان می‌توان به روشهای مبتنی بر افراز داده‌ها[۲]، روشهای سلسله مراتبی[۳]، روشهای شبکه‌ای[۴]، روشهای مبتنی بر تراکم[۵] و روشهای مبتنی بر مدل[۶] اشاره نمود [۱و۲].

از میان روشهای بسیار زیادی که برای خوشه‌بندی داده‌ها طراحی و ارائه شده است، اغلب آنها تنها به منظور خوشه‌بندی نوع خاصی از داده‌ها (داده‌هایی که صرفا از نوع عددی یا صرفا از نوع دسته‌ای[۷] باشند) طراحی شده‌اند و بر روی داده‌های مختلط که انواع مختلف مشخصه‌های عددی و دسته ای را شامل می‌شوند، قابلیت کاربرد ندارند. از سوی دیگر بسیاری از مسایل دنیای واقعی و اغلب پایگاههای داده‌ای که روشهای خوشه‌بندی به منظور تحلیل آنها ایجاد شده‌اند، داده هایی از نوع مختلط را در خود جای داده‌اند. بنابراین روش خوشه‌بندی که قابلیت کار بر

تاریخ وصول: ۸۹/۳/۸
تاریخ تصویب: ۹۰/۴/۵

***نویسنده مسئول مقاله: دکتر مسعود یقینی**، استادیار، دانشکده مهندسی راه آهن، دانشگاه علم و صنعت ایران، *yaghini@iust.ac.ir*
مهدی ورد، کارشناسی ارشد، دانشکده مهندسی راه آهن، دانشگاه علم و صنعت ایران، *Mahdivard@yahoo.com*

[2] *Partitioning Methods*
[3] *Hierarchical Methods*
[4] *Grid-based Methods*
[5] *Density-based Methods*
[6] *Model-based Methods*
[7] *Categorical*

روی داده‌های مختلط را دارا باشد، بسیار مورد توجه و حائز اهمیت خواهد بود و چنین روشی می‌تواند در حل بسیاری از مسایل دنیای واقعی و تحلیل پایگاههای داده در سازمانهای مختلف مورد استفاده قرار گیرد.

برای طراحی روشی که بتواند علاوه بر داده‌های عددی، بر روی داده‌های دسته‌ای نیز قابلیت کاربرد داشته باشد، نخستین موضوعی که باید به آن پرداخته شود طراحی معیار سنجش فاصله میان داده‌های دسته‌ای است. در این زمینه روشهای مختلفی ارائه شده است که روش هوآنگ [۳] با توجه به سادگی و کاربرد آسان، از روشهای مورد توجه است. اما روش هوآنگ علیرغم سادگی، از دقت خوبی برخوردار نبوده و دارای نواقصی است که سبب می‌شود عملکرد الگوریتم خوشه‌بندی با این معیار فاصله از کیفیت خوبی برخوردار نباشد. لذا در این مقاله سعی نموده‌ایم با بهره گیری از یک معیار فاصله دقیق‌تر، الگوریتمی را طراحی کنیم که از دقت بالاتری نسبت به دیگر روشهای خوشه‌بندی داده‌های مختلط برخوردار باشد.

یکی دیگر از بخشهای تشکیل‌دهنده یک روش خوشه‌بندی، انتخاب رویکردی جهت جستجوی فضای جواب و بهینه‌سازی تابع هدف است. یکی از تکنیکهای جستجوی فضای جواب و بهینه-یابی، روشهای فرا ابتکاری هستند که از آن جمله می‌توان به الگوریتم ژنتیک [۴]، الگوریتم کلونی مورچگان [۵] و روش جستجوی ممنوع [۶] اشاره نمود.

از میان روشهای فرا ابتکاری، 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 خوشه‌بندی نیز از این روش به عنوان ابزار مورد استفاده جهت بهینه‌سازی تابع هدف استفاده نموده‌اند. لذا با توجه به قابلیت و کارایی بالای الگوریتم ژنتیک، ما نیز در طراحی روش خوشه‌بندی خود از این الگوریتم جهت جستجوی فضای جواب بهره برده‌ایم.

یکی از ویژگیهای روش خوشه‌بندی ارائه شده در این مقاله، محاسبه تعداد بهینه خوشه‌هاست؛ به این معنی که بر خلاف بسیاری از روشهای خوشه‌بندی که تعداد خوشه‌ها را به عنوان یک مقدار ورودی از کاربر دریافت می‌کنند، روش پیشنهادی ما می-تواند به موازات گروه‌بندی داده‌ها و تشکیل خوشه‌های بهینه، مقدار بهینه برای تعداد خوشه‌ها را نیز محاسبه و ارائه دهد.

در ادامه ابتدا مروری بر روشهای خوشه‌بندی مبتنی بر الگوریتم ژنتیک و نیز روشهای خوشه‌بندی داده‌های مختلط صورت گرفته است. سپس به تشریح روش مورد استفاده برای محاسبه فاصله بین مقادیر داده های دسته‌ای خواهیم پرداخت و پس از آن مراحل الگوریتم ژنتیک مورد استفاده در فرآیند خوشه‌بندی معرفی خواهد شد.

۲. مروری بر ادبیات موضوع

۲-۱. روشهای خوشه‌بندی مبتنی بر الگوریتم ژنتیک

کریشنا و مورتی [۷] یک روش خوشه‌بندی مبتنی بر الگوریتم ژنتیک (GKA) ارائه کردند که می‌توان برای حل مساله خوشه-بندی با تعداد خوشه‌های مشخص از آن بهره گرفت.

لو و همکاران [۸] الگوریتم سریع k-means مبتنی بر ژنتیک (FGKA)[1] را با الهام از الگوریتم GKA ابداع نمودند که این روش در بسیاری از جنبه‌ها نسبت به GKA بهبود داده شده است. از جمله اینکه روش FGKA بسیار سریع‌تر از GKA عمل می‌کند.

لو و همکاران [۹] الگوریتم افزاینده k-means ژنتیک[2] (IGKA) را طراحی کردند که توسعه‌ای بر الگوریتم خوشه‌بندی قبلی آنها (روش FGKA) بود. روش IGKA این ویژگی برجسته را از روش FGKA به ارث برده است که همواره به جواب بهینه کلی همگرا می‌باشد.

ماولیک و همکاران [۱۰] یک روش خوشه‌بندی 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 می‌شود. در این روش مقدار K از پیش معلوم در نظر گرفته شده است و یکی از مقادیر ورودی مساله می باشد.

در روش دیگری که توسط لین و همکاران [۱۱] ارائه گردید، مراکز خوشه ها مستقیما از میان نقاط مجموعه داده انتخاب می-شود. اتخاذ این رویکرد سبب سرعت بیشتر در محاسبه تابع برازندگی و در نتیجه سریع تر شدن کل الگوریتم می گردد.

بندیوپد و ماولیک [۱۲] روشی را ارائه دادند که در آن همزمان با خوشه بندی داده ها، مقدار مناسب برای تعداد خوشه ها نیز تعیین می شود. در این روش جهت اعتبار سنجی خوشه های حاصله از شاخص Davies–Bouldin استفاده شده است.

یک روش خوشه بندی ترکیبی مبتنی بر الگوریتم ژنتیک تحت عنوان روش خوشه بندی HGA توسط لیو و همکاران [۱۳] ارائه شد. این الگوریتم با بهره گیری از لیست ممنوعه و معیار انتظار، بین تنوع در جمعیت و سرعت همگرایی، هماهنگی ایجاد می کند.

چیانگ و همکاران [۱۴]، روش k-modes را که با ایجاد تغییراتی در روش k-means، به خوشه بندی داده‌های دسته ای می پردازد، توسعه دادند. به این ترتیب که با بهره‌گیری از الگوریتم ژنتیک، شاخصی برای سنجش عدم تشابه با نام مقیاس فاصله ژنتیک (GDM) طراحی نمودند.

[1] *Fast Genetic K-means Algorithm*
[2] *Incremental Genetic K-means Algorithm*

الگوریتم k-means ژنتیک وزن‌دار[1] (GWKMA) که تلفیقی از الگوریتم ژنتیک و الگوریتم k-means وزن دار است، توسط ژیانگ و همکاران [۱۵] پیشنهاد گردید.

HGACLUS، مدلی ترکیبی جهت خوشه بندی بر مبنای الگوریتم ژنتیک است که توسط پن و ژو [۱۶] ارائه گردید. در این مدل از روش Simulated Annealing برای یافتن مراکز خوشه ها استفاده شده است.

کاتاری و همکاران [۱۷] ، روشی را برای خوشه بندی داده ها ارائه نمودند که در آن از الگوریتم ژنتیک بهبود یافته[2] استفاده شده و عملگرهای باز ترکیب و جهش به شکل کاراتری تعریف شده اند.

۲-۲. روشهای خوشه بندی داده های مختلط

امروزه با پایگاههای داده بسیار بزرگی مواجه هستیم که مشتمل بر مشخصه‌های مختلط هستند. در حالیکه اغلب روشهای خوشه‌بندی که تا کنون ارائه شده اند تنها بر روی داده های عددی و یا دسته ای عملکرد خوبی داشته و قابلیت کاربرد بر روی داده هایی که دارای مشخصه هایی از هر دو نوع عددی و دسته ای هستند را ندارند.

برای غلبه بر این مشکل، برخی از استراتژی های به کار گرفته شده به شرح زیر می باشند [۱۸]:

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 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 گسسته‌سازی مقادیر عددی با از دست دادن اطلاعات توام است.

لی و بیزواز [۱۹] یک الگوریتم خوشه بندی ترکیب کننده مبتنی بر تشابهات[3] را ارائه کردند که بر اساس شاخص تشابه گودال [۲۰] 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 [۳] تابع هزینه ای پیشنهاد کرد که مشخصه های عددی و دسته ای را به صورت مجزا در نظر می گیرد. این تابع هزینه در الگوریتم‌های خوشه‌بندی مبتنی بر افراز داده‌ها 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 داده‌ها را به صورت مجموع دو مقدار فاصله، یکی برای مشخصه‌های عددی و دیگری برای مشخصه های دسته ای، محاسبه می‌نماید. از آنجا که تابع هزینه هوآنگ قابلیت کاربرد با الگوریتم های مبتنی بر افراز را دارد، هزینه های محاسباتی آن در حد مناسب و قابل قبولی است.

پس از آن، هوآنگ و همکاران [۲۱] 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 همکاران [۲۲] روشی را پیشنهاد کردند که در آن مشخصه های عددی و دسته ای به صورت جداگانه خوشه بندی می‌شوند و از تکنیک جمع آوری شواهد برای ترکیب نتایج خوشه بندی ها و حصول خوشه بندی نهایی استفاده می شود.

هی و همکاران [۲۳]، روش قبلی خود را که به خوشه‌بندی داده‌های دسته ای می پرداخت و الگوریتم فشرده[4] نام داشت توسعه دادند و روشی را پیشنهاد کردند که قابلیت کاربرد بر روی داده های مختلط را دارد.

احمد و دی [۱۸] تابع هزینه ای را ارائه دادند که در آن سعی شده با بهبود و رفع نقایص تابع هوآنگ، فرآیند خوشه‌بندی 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 دسته‌ای، براساس 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 مشخصه‌ها برای دو شیء مورد بررسی، مقداری بین صفر و یک را محاسبه و آنرا به عنوان فاصله دو شیء در نظر می‌گیرد.

۳. معرفی روش خوشه‌بندی پیشنهادی

روش خوشه‌بندی پیشنهادی در این مقاله از نوع روشهای افراز داده‌ها و مبتنی بر الگوریتم ژنتیک است که در آن از الگوریتم ژنتیک جهت جستجوی جواب بهینه استفاده می‌شود. همچنین این روش قادر به خوشه‌بندی داده‌های مختلط می‌باشد. به طور کلی مهم‌ترین مشخصات روش پیشنهادی عبارتند از:

[1] *Genetic Weighted K-means Algorithm*
[2] *Improved Genetic Algorithm (IGA)*
[3] *Similarity Based Agglomerative Clustering (SBAC)*
[4] *Squeezed Algorithm*

۱) قابلیت کار بر روی داده‌های مختلط
۲) تعیین مقدار بهینه برای تعداد خوشه‌ها
۳) استفاده از روش دقیق‌تری برای تعیین فاصله بین داده‌های دسته‌ای
۴) استفاده از شاخص Davies-Bouldin به عنوان تابع برازندگی
نوآوری مقاله حاضر، ارائه الگوریتمی است که مجموعه ویژگیهای فوق را به صورت توام دارا می‌باشد.

۳-۱. روش مورد استفاده برای محاسبه فاصله بین دو مقدار از یک متغیر دسته‌ای

در روش خوشه‌بندی پیشنهادی در این مقاله، از تابع فاصله ارائه شده توسط احمد و دی [۱۸] جهت محاسبه فاصله میان نقاط و نیز محاسبه مراکز خوشه‌ها استفاده شده است و جهت پیاده‌سازی الگوریتم ژنتیک، نحوه نمایش جوابها در کروموزومها و نیز عملگرهای الگوریتم ژنتیک متناسب با این تابع فاصله تعریف شده است.

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:

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

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

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

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

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

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

۳-۲. تعیین فاصله میان دو داده

فرض کنید که D_1 و D_2 دو داده از مجموعه داده های مختلط باشند که مجموعا دارای m مشخصه می باشند که m_r مشخصه اول عددی و m_c مشخصه بعدی دسته‌ای هستند و $m_r + m_c = m$ فاصله بین D_1 و D_2 برابر است با:

$$Dist(D_1, D_2) = \sum_{t=1}^{m_r} (w_t(X_t - Y_t))^2 + \sum_{t=1}^{m_c} (\delta(X_t, Y_t))^2 \qquad (۴)$$

۳-۳. محاسبه مراکز خوشه ها برای داده های 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 این مقاله از این تعریف برای مراکز خوشه ها در حالت خوشه بندی با مرز مشخص[1] 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 مشخصه‌های دسته‌ای از نحوه نمایش متفاوتی استفاده شده است. از آنجا که در روش پیشنهادی فاصله بین دو مقدار دسته‌ای براساس توزیع کلی آنها در سراسر مجموعه داده‌ها تعریف می شود، این مقدار فاصله به ازای زوجهای مختلف از مقادیر، متفاوت خواهد بود.
بنابراین اگر به عنوان مثال فاصله مقدار r تا مقدار s کمتر از فاصله r تا t باشد، یعنی $\delta(r,s) < \delta(r,t)$ آنگاه انتظار می رود که در یک خوشه بندی مناسب از داده ها، تعداد رخداد های همزمان[2] r و s از تعداد رخداد های همزمان r و t بیشتر باشد.
با در نظر گرفتن این مطالب، مقدار مرکزی aامین مشخصه دسته‌ای برای خوشه C به شکل زیر محاسبه می گردد:

$$1/N_c \langle (N_{1,1,,c}, N_{1,2,,c}, \ldots, N_{1,p1,c}), \ldots, (N_{m,1,c}, N_{m,2,c}, \ldots, N_{m,pm,c}) \rangle \qquad (۵)$$

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 دسته‌ای را در خوشه مورد نظر نشان می دهد.

[1] *Crisp Clustering*
[2] *Co-occurrence*

۳-۴. 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 دسته‌ای، 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 نمایانگر نسبت حضور مقدار v در خوشه است، تخصیص داده می شود. فرض کنید $a_{i,k}$ نمایانگر kامین مقدار ممکن برای مشخصه دسته ای 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:

$$\Omega(X,C)=\left(N_{i,1,c}/N_c\right)*\delta\left(X,A_{i,1}\right)+\ldots +\left(N_{i,p_i,c}/N_c\right)*\delta\left(X,A_{i,p_i}\right) \quad (۶)$$

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

$$D(d_i,C_j)=\sum_{t=1}^{m_r}\left(d_{it}^r-C_{jt}^r\right)^2+\sum_{t=1}^{m_c}\left(\Omega(d_{it}^c,C_{jt}^c)\right)^2 \quad (۷)$$

۳-۵. اجزاء الگوریتم ژنتیک در روش پیشنهادی

۳-۵-۱. نحوه نمایش رشته ها

در روش پیشنهادی، کروموزوم ها از اعداد حقیقی تشکیل شده‌اند و مقادیر و مختصات مربوط به مراکز خوشه‌ها را در خود جای داده اند. طول کروموزوم‌ها 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 می‌بایست 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 مقدار ۱-) تخصیص داده می شود تا مشخص شود که ژن مربوطه خالی است و مرکز خوشه ای در آن قرار نگرفته است.

۳-۵-۲. مقدار دهی اولیه جمعیت

به ازای هر کروموزوم i در جمعیت ($i=1, \ldots, P$ و P برابر با اندازه جمعیت است)، یک مقدار تصادفی k_i در بازه تعریف شده تولید می‌شود. سپس k_i نقطه به صورت تصادفی از میان داده‌ها انتخاب می شود و به صورت تصادفی در میان ژنهای کروموزوم قرار داده می‌شود. در نهایت به ژنهای خالی کروموزوم مقدار (۱-) تخصیص داده می شود.

۳-۵-۳. عملگرهای تغییر

در روش پیشنهادی از عملگر باز ترکیب تک نقطه‌ای استفاده شده است. مقدار p_c در آزمایشهای مختلف بین ۰/۵ تا ۰/۷ قرار داده شد و نتایج حاصله مقایسه شد و در نهایت نرخ بازترکیب برابر ۰/۵ انتخاب گردید. در مورد عملگر جهش نیز یک تعریف جدید برای اعمال عملگر جهش بر روی مقادیر دسته‌ای ابداع شد. همانطور که در بخش قبلی اشاره شد، نمایش مختصات مراکز خوشه ها در مورد مشخصه‌های دسته‌ای به صورت نسبت فراوانی هر یک از مقادیر مشخصه مزبور در میان نقاط موجود در خوشه مورد نظر است. برای اجرای عملگر جهش، در صورت انتخاب مشخصه دسته‌ای A_i (احتمال انتخاب هر یک از مشخصه‌ها برای اجرای عملگر جهش برای با مقدار p_m است) از ردیف مربوط به مشخصه A_i در ماتریس نمایش مختصات مرکز خوشه ها، دو نسبت a_{ij} و a_{ik} انتخاب می شود و مقادیر آنها با هم عوض می-شود.

۳-۵-۴. تابع برازندگی

یکی دیگر از مواردی که در طراحی الگوریتم‌های خوشه‌بندی باید مدنظر قرار گیرد، انتخاب مقیاس اعتبار مناسب جهت انتخاب به عنوان تابع برازندگی است. شاخصهای اعتبار مختلفی نظیر شاخص Dunn، شاخص XB (Xie-Beni)، شاخص BM و شاخص DB در این زمینه ارائه شده‌اند. شاخص DB، که به صورت تابعی از نسبت مجموع پراکندگی نقاط در داخل خوشه به جدایی بین خوشه‌ها تعریف می‌شود، در مقایسه با سایر شاخصهایی که در بالا به آنها اشاره شد نتایج دقیق تری را به دست می‌دهد. در روش ارائه شده در این مقاله از شاخص DB به عنوان تابع برازندگی استفاده شده است. مقادیر کوچکتر این شاخص نشاندهنده خوشه بندی داده ها به نحوی بهتر خواهد بود. با توجه به اینکه فرآیند الگوریتم ژنتیک به دنبال بیشینه سازی

تابع هدف است، معکوس این شاخص به عنوان مقدار تابع برازنگی تعریف شده است. روابط مربوط به محاسبه شاخص DB در ادامه ارائه می‌گردد.

$$S_i = \frac{1}{|C_i|}\sum_{x\in C_i}\|x - z_i\| \quad (۸)$$

$$R_i = \underset{j,j\neq i}{Max}\left\{\frac{S_i + S_j}{d_{ij}}\right\} \quad (۹)$$

$$d_{ij} = d(C_i, C_j) = \|z_i - z_j\| \quad (۱۰)$$

$$DB_r = \frac{1}{k_r}\sum_{i=1}^{k_r} R_i \quad (۱۱)$$

$$Fitness(Ch_r) = \frac{1}{DB_r} \quad (۱۲)$$

که در روابط فوق C_i نمایانگر خوشه iام، $|C_i|$ بیانگر تعداد نقاط موجود در خوشه iام، z_i نشاندهنده مرکز خوشه iام و DB_r نمایانگر مقدار شاخص DB برای کروموزوم r است.

۴. آزمایش الگوریتم پیشنهادی

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

۴-۱. تنظیم پارامترهای ورودی الگوریتم

با استفاده از مسایل شبیه سازی شده، مقادیر پارامترهای ورودی الگوریتم شامل تعداد جمعیت، نرخ باز ترکیب، نرخ جهش، حداکثر تعداد نسلها و نیز k_{min} و k_{max} تعیین گردید که مقادیر این پارامترها در جدول ۱ آورده شده است. همچنین در الگوریتم پیشنهادی، انتخاب والدها بر اساس روش متناسب با برازندگی و انتخاب بازمانده‌ها بر اساس سن کروموزوم‌ها صورت می‌گیرد.

۴-۲. نتایج الگوریتم بر روی داده های استاندارد

در این قسمت نتایج اجرای الگوریتم خوشه بندی پیشنهادی بر روی مجموعه داده‌های استاندارد آورده شده است. به منظور مقایسه نحوه عملکرد الگوریتم ارائه شده با روشهای قبلی، از دو مجموعه داده استاندارد استفاده شده است که بسیاری از مقالات معتبر برای آزمایش الگوریتم خود از آن استفاده کرده اند و لذا امکان مقایسه نتایج وجود خواهد داشت. این داده ها از مخزن داده UCI[1] استخراج شده اند. مجموعه داده‌های مورد استفاده به صورت از پیش طبقه بندی شده هستند و کلاس متناظر با هر داده از پیش معلوم است. در نتیجه برای سنجش میزان دقت الگوریتم، از میزان انطباق نحوه خوشه بندی داده ها با کلاسهای واقعی آنها استفاده گردیده است. در ادامه نتایج اجرای الگوریتم بر روی هر یک از این دو مجموعه داده استاندارد آورده شده است.

جدول ۱. مقادیر پارامترهای ورودی الگوریتم پیشنهادی

پارامتر ورودی	مقدار پارامتر
حداقل تعداد خوشه ها (k_{min})	2
حداکثر تعداد خوشه ها (k_{max})	15
حداکثر تعداد نسلها (max-gen)	25
نرخ باز ترکیب (p_c)	0.5
نرخ جهش (p_m)	0.01
اندازه جمعیت	40

۴-۲-۱. داده های بیماران قلبی[2]

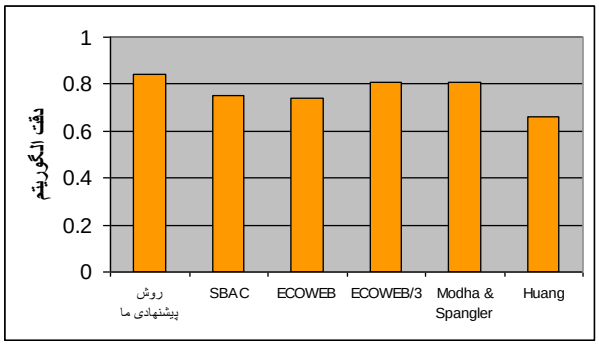
این داده ها اطلاعات مربوط به تعدادی از بیماران قلبی را شامل می شود و در کلینیک کلولند تولید شده است. پایگاه داده اصلی شامل ۷۶ مشخصه است، اما مقالات تحقیقی مختلف برای آزمایش الگوریتم خود از یک زیر مجموعه از این پایگاه داده که شامل ۱۴ مشخصه است استفاده نموده اند. مجموعه داده های بیماران قلبی یک مجموعه داده مختلط است که شامل ۹ مشخصه دسته ای و ۵ مشخصه عددی می‌باشد. 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 ۲ نتایج حاصل از خوشه بندی این مجموعه داده را به وسیله روش خوشه بندی پیشنهادی در این مقاله نشان می‌دهد. 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 [۱۹] روش ECOWEB [۲۴]، روش COBWEB/3 [۲۵]، روش مودها و اسپنگلر [۲۶] و روش هوآنگ [۳] آورده شده است. مقادیر دقت بدست آمده برای این الگوریتمها، دقت بیشتر و برتری روش خوشه بندی ارائه شده را نسبت به سایر روشها نشان می‌دهد.

[1] *http://archive.ics.uci.edu*
[2] *Heart Disease Data*

جدول ۲. مقایسه نتایج روشهای مختلف بر روی مجموعه داده بیماران قلبی

نام الگوریتم	تعداد داده هایی که در خوشه مورد انتظار قرار گرفته‌اند	دقت
روش پیشنهادی	255	0.84
SBAC	228	0.75
ECOWEB	224	0.74
COBWEB/3	245	0.81
روش مودها و اسپنگلر	244	0.81
روش هوآنگ	200	0.66



شکل ۱: مقایسه میزان دقت روش خوشه بندی پیشنهادی نسبت به دیگر روشها بر روی داده های بیماران قلبی

۲-۲-۴. داده های کارتهای اعتباری[1]

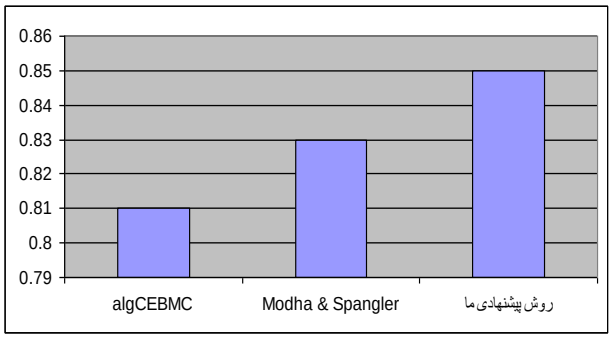
این مجموعه داده، اطلاعات مربوط به یک موسسه مالی و اعتباری در استرالیا را شامل می‌شود. این داده ها یک مجموعه داده مختلط هستند که دارای هشت مشخصه دسته‌ای و شش مشخصه عددی می باشند. این مجموعه داده دارای ۶۹۰ نمونه است که به دو کلاس تقسیم می‌شوند: کلاس منفی شامل ۳۸۳ نمونه و کلاس مثبت شامل ۳۰۷ نمونه.

نتایج حاصل از خوشه بندی این مجموعه داده نیز توسط روش پیشنهادی و دو روش خوشه بندی داده های مختلط که برای آزمایش نتایج خود از مجموعه داده های کارتهای اعتباری استفاده کرده اند، در جدول ۳ آورده شده است. مقادیر این جدول نیز بار دیگر برتری روش پیشنهادی را نسبت به روش ارائه شده توسط مودها و اسپنگلر [۲۶] تائید می کند.

[1] *Australian Credit Approval*

جدول ۳. مقایسه نتایج روشهای مختلف بر روی مجموعه داده کارتهای اعتباری

نام الگوریتم	تعداد داده هایی که در خوشه مورد انتظار قرار گرفته‌اند	دقت
روش پیشنهادی	587	0.85
روش مودها و اسپنگلر	572	0.83
algCEBMC	559	0.81



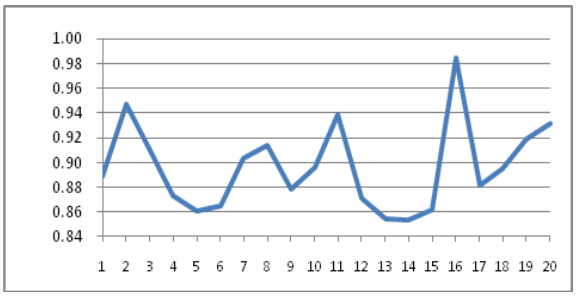
شکل ۲. مقایسه میزان دقت روش خوشه بندی پیشنهادی نسبت به دیگر روشها بر روی داده های موسسه اعتباری

۳-۴. نتایج الگوریتم پیشنهادی بر روی داده های شبیه-سازی شده

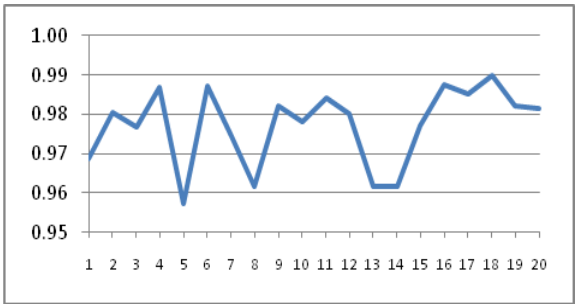
در این بخش جهت سنجش نتایج الگوریتم پیشنهادی، مجموعه داده‌هایی با ابعاد ۲۰۰۰، ۲۰۰۰۰ و ۱۲۰۰۰۰ داده شبیه سازی شده و فرآیند خوشه‌بندی در مورد آنها انجام گرفت.

همچنین با تغییر نحوه تعریف فاصله بین مقادیر دسته ای در روش خوشه‌بندی 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-prototypeS بهبود یافته با دقت بیشتری صورت گیرد، از دو شاخص اعتبار سنجی یعنی شاخص مجموع مربعات خطاها (SSE) و شاخص DB استفاده گردید و مقادیر هر دوی این شاخصها برای جواب حاصل از هر یک از دو روش خوشه‌بندی محاسبه شد. شکلهای ۳ تا 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 در نمودارهای فوق الذکر قابل مشاهده است، اینست که علیرغم اینکه در روش پیشنهادی از شاخص DB به عنوان تابع برازندگی استفاده شده و الگوریتم پیشنهادی به دنبال کمینه کردن این تابع هزینه است،

اما به ازای تمامی مجموعه داده‌های شبیه سازی شده، مقدار شاخص SSE نیز برای الگوریتم پیشنهادی مقدار کمتری را به خود اختصاص داده است.

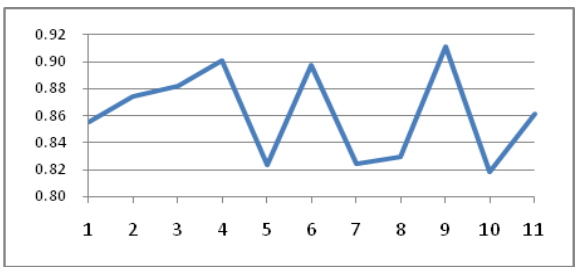


شکل ۳. نسبت مقدار شاخص DB برای روش پیشنهادی به مقدار این شاخص برای روش k-prototypes بهبود یافته؛ مجموعه داده‌های ۲۰۰۰ تایی

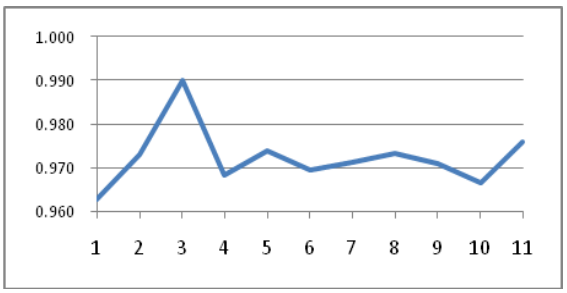


شکل ۴. نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش k-prototypes بهبود یافته؛ مجموعه داده‌های ۲۰۰۰ تایی

شکلهای ۳ و ۴ نشان می‌دهد که در مورد مجموعه داده‌های ۲۰۰۰ تایی، مقدار شاخص DB برای روش پیشنهادی بین ۲٪ تا ۱۴٪ نسبت به مقدار این شاخص برای روش k-prototypes کمتر است. شاخص SSE روش پیشنهادی نیز بین ۱٪ تا ۴٪ نسبت به شاخص SSE روش k-prototypes کمتر است.

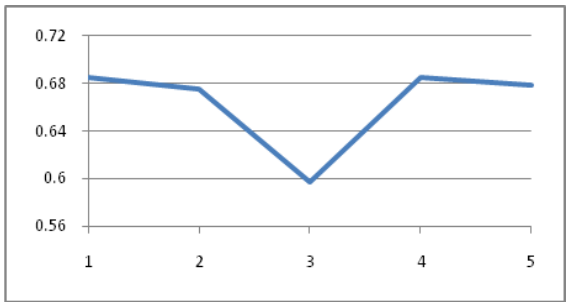


شکل ۵. نسبت مقدار شاخص DB برای روش پیشنهادی به مقدار این شاخص برای روش k-prototypes بهبود یافته؛ مجموعه داده‌های ۲۰۰۰۰ تایی

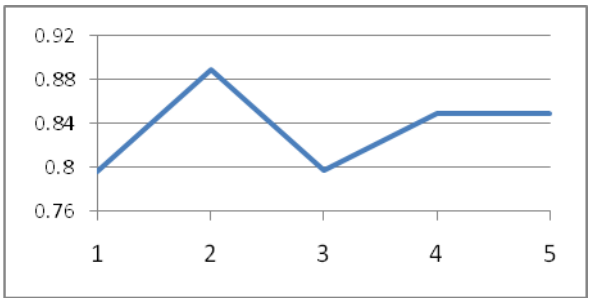


شکل ۶. نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش k-prototypes بهبود یافته؛ مجموعه داده های ۲۰۰۰۰ تایی

شکلهای ۵ و ۶ بیانگر برتری نتایج حاصل از روش پیشنهادی نسبت به روش k-prototypes در مورد مجموعه داده‌های ۲۰۰۰۰ تایی است؛ به نحویکه مقدار شاخص DB برای روش پیشنهادی بین ۹٪ تا ۱۸٪ و مقدار شاخص SSE برای روش پیشنهادی بین ۱٪ تا ۴٪ نسبت به شاخصهای متناظر برای روش k-prototypes کمتر است.



شکل ۷. نسبت مقدار شاخص DB برای روش پیشنهادی به مقدار این شاخص برای روش k-prototypes بهبود یافته؛ مجموعه داده‌های ۱۲۰۰۰۰ تایی



شکل ۸: نسبت مقدار شاخص SSE برای روش پیشنهادی به مقدار این شاخص برای روش k-prototypes بهبود یافته؛ مجموعه داده‌های ۱۲۰۰۰۰ تایی

شکلهای ۷ و ۸ نیز بار دیگر برتری روش پیشنهادی را در مورد هر دو معیار DB و SSE تایید می‌نماید. به نحویکه مقدار شاخص

DB برای روش پیشنهادی بین ۳۱٪ تا ۴۰٪ و مقدار شاخص SSE برای روش پیشنهادی بین ۱۱٪ تا ۲۰٪ نسبت به شاخصهای متناظر برای روش k-prototypes کمتر است.

۵. نتیجه‌گیری

در این مقاله یک روش خوشه‌بندی داده‌های مختلط مبتنی بر الگوریتم ژنتیک ارائه شده است. در روش پیشنهادی، برخلاف اکثر روشهای خوشه‌بندی داده‌های مختلط که از معیار فاصله صفر و یک برای اندازه‌گیری فاصله بین داده‌های دسته‌ای بهره می‌برند، از تعریف دقیق‌تری جهت سنجش فاصله بین داده‌های دسته‌ای و نیز محاسبه مراکز خوشه‌ها استفاده شده است و سپس اجزاء الگوریتم ژنتیک (عملگرهای بازترکیب و جهش) متناسب با ساختار جدید نمایش مراکز خوشه‌ها برای هر یک از داده‌های عددی و دسته‌ای تعریف گردیده است.

از دیگر مزایای روش پیشنهادی اینست که نیازی به تعیین تعداد خوشه‌ها به عنوان ورودی الگوریتم نداشته و قادر است با بهره گیری از قابلیت جستجوی الگوریتم ژنتیک در فضای جواب، ضمن خوشه‌بندی داده‌ها، مقدار بهینه تعداد خوشه‌ها را نیز محاسبه نماید که این ویژگی در مورد بسیاری از مسایل دنیای واقعی که در آنها با مجموعه داده‌های بسیار بزرگ با تعداد خوشه‌های نامعین سر و کار داریم، بسیار حائز اهمیت است. آزمایش الگوریتم پیشنهادی توسط داده‌های استاندارد و نیز داده‌های شبیه‌سازی شده، نشان از برتری این روش و دقت بالاتر آن نسبت به سایر روشهای خوشه‌بندی داده‌های مختلط دارد.

مراجع

[1] Han, J., Kamber, M., *Data Mining Concepts And Techniques*, Elsevier, 2006.

[2] Witten, L.H., Frank, E., *Data Mining-Practical Machine Learning Tools And Techniques*, Elsevier, 2005.

[3] Huang, Z., *"Clustering Large Data Sets with Mixed Numeric & Categorical Values"*, in: Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining, World Scientific, Singapore, 1997.

[4] Goldberg, D.E., *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, 1989.

[5] Dorigo, M., Stützle, T., *Ant Colony Optimization*, Cambridge: MIT Press, 2004.

[6] Glover, F., Laguna, M., *Tabu search*, Kluwer Academic Publishers, Boston, 1997.

[7] Krishna, K., Murty, M.N., *"Genetic K-Means Algorithm"*, IEEE Transaction On Systems, Man, And cybernetics—Part B:CYBERNETICS, Vol. 29, No. 3, 1999, pp. 433-439.

[8] Lu, Y., Lu, S., Fotouhi, F., "*FGKA: A Fast Genetic K-means Clustering Algorithm*", SAC'04 Nicosia, Cyprus. , ACM, 2004.

[9] Lu, Y., Lu, S., Fotouhi, F., Deng, Y., Susan, D., Brown, J., "*an Incremental Genetic K-Means Algorithm and its Application in Gene Expression Data Analysis*", BMCBioinformatics, Vol. 5, 2004, pp. 172-181.

[10] Maulik, U., Bandyopadhyay, S., "*Genetic Algorithm-Based Clustering Technique*", Pattern Recognition, Vol. 33, No. 9, 2000, pp. 1455-1465.

[11] Lin, H.J., Yang, F.W., Kao, Y.T., "*An Efficient GA Based Clustering Technique*", Tamkang Journal of Science and Engineering, Vol. 8, No. 2, 2005, pp. 113- 122.

[12] Bandyopadhyay, S., Maulik, U., "*Genetic Clustering for Automatic Evolution of Clusters and Application to Image Classification*", PatternRecognition,Vol.35, No. 6, 2002, pp. 1197- 1208.

[13] Liu, Y., Kefe, S., Liz, X., "*A Hybrid Genetic Based Clustering Algorithm*", Proceedings of the Third International Conference on Machine Leaming and Cybernetics, Shanghai, 2004.

[14] Chiang, S., Chu, S.C., Hsin, Y.C., Wang, M.H., "*Genetic Distance Measure for K-Modes Algorithm*", International Journal of Innovative Computing, Information and Control, Vol. 2, No. 1, 2006, pp. 33-40.

[15] Wu, F.X., Kusalik, A.J., Zhang, W.J., "*Genetic Weighted K-means for Large-Scale Clustering Problems*", University of Saskatchewan, CANADA, 2003.

[16] Pan, H., Zhu, J., *"Genetic Algorithms Applied to Multi-Class Clustering for Gene Expression Data"*, Genomics Proteomics Bioinformatics, Vol. 1, No. 4, 2003, pp. 279-287.

[17] Katari, V., Satapathy, S.C., Murthy, J., Reddy, P., *"Hybridized Improved Genetic Algorithm with*

Variable Length Chromosome for Image Clustering", International Journal of Computer Science and Network Security, Vol. 7, No. 11, 2007, pp. 121-131.

[18] Ahmad, A., Dey, L., "*A k-Mean Clustering Algorithm for Mixed Numeric and Categorical Data*", Data & Knowledge Engineering, Vol. 63, No. 2, 2007, pp. 503- 527.

[19] Li, C., Biswas, G., *"Unsupervised Learning with Mixed Numeric and Nominal Data"*, IEEE Transactions on Knowledge and Data Engineering Vol. 14, No. 4, 2002, pp. 673- 690.

[20] Goodall, D.W., *"A New Similarity Index Based on Probability"*, Biometric 22, 1966, pp. 882- 907.

[21] Huang, J.Z., Ng, M.K., Rong, H., Li, Z., "*Automated Variable Weighting in k-Mean Type Clustering*", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 5, 2005, pp. 657- 668.

[22] Luo, H., Kong, F., Li, Y., "*Clustering Mixed Data Based on Evidence Accumulation*", Lecture Notes on Artificial Intelligence 4093, 2006.

[23] He, Z., Xu, X., Deng, S., *"Scalable Algorithms for Clustering Large Datasets with Mixed Type Attributes"*, International Journal of Intelligent Systems, Vol 20, No. 10, 2005, pp. 1077- 1089.

[24] Reich, Y., Fenves, S.J., "*The Formation and use of Abstract Concepts in Design*", Morgan Kaufmann Series In Machine Learning, 1991, pp. 323- 353.

[25] McKusick, K., Thompson, K., "*COBWEB/3: A Portable Implementation*", Technical Report FIA-90-6-18-2, NASA Ames Research Center, 1990.

[26] Modha, D.S., Spangler, W.S., *"Feature Weighting in k-Mean Clustering"*, Machine Learning, Vol. 52, No. 3, pp. 217- 237.

[27] 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, 2001.