فصلنامه پژوهش در علوم، مهندسی و فناوری
دوره ۴، شماره ۲، تابستان ۱۳۹۷
صفحات ۶۵-۸۴
ISSN: 2476-5082
www.uctjournals.com

نیم قرن پس از خوشه بندی؛ بررسی و ارزیابی رویکرد ها و روشهای خوشه بندی با تجزیه و تحلیل تصمیم گیری چند معیاره

عباس سرافرازی[۱]

[۱] دانشجوی دکتری مهندسی صنایع و مربی، دانشگاه پیام نور، تهران، ایران

چکیده

امروزه خوشه‌بندی به عنوان یک روش یادگیری بدون ناظر در کاربردهای بسیاری توانسته است ارزش خود را نشان دهد. یکی از روشهای حیاتی کنترل و مدیریت داده ها، کلاس بندی یا گروه بندی داده هایی با خواص مشابه درون مجموعه ای از دسته ها یا خوشه ها می باشد. برای این منظور لازم است که الگوهایی با بیشترین میزان شباهت، در یک خوشه قرار گیرند. رویکردهای اصلی خوشه بندی عبارتند از: افرازی، سلسله مراتبی، مبتنی بر چگالی، مبتنی بر مشبک کردن فضا، نقشه های خود سازمانده، متاهیورستیک. در این پروژه از مدلهای چند شاخصه استفاده گردیده که بطور کلی جهت انتخاب مناسبترین گزینه از بین m گزینه موجود می باشد لذا تصمیم گیریهای چند شاخصه بصورت ماتریسی نمایش داده می شود که تعداد سطرهای مبین گزینه های موجود و تعداد ستونها مبین شاخصهای و معیارهای موجود می باشد. سپس جهت اولویت بندی گزینه های موجود از روش تاپسیس استفاده گردید. با توجه به نظر سنجی از خبرگان, متوسط نظرات خبرگان در جدول تصمیم ماتریسی با ابعاد ۱۹*۷ در خصوص بررسی رویکردهای و ۱۹*۲۱ درباره روشهای خوشه بندی تولید شد. سطر های ماتریس شامل انواع رویکرهای مختلف به تحلیل خوشه ای داده ها می باشد و ستونهای ماتریس نیز شامل ۱۹ معیاری است که بر اساس آن معیارها مناسب ترین رویکرد و روش تجزیه و تحلیل از منظر تست تئوری بررسی می شود. نتایج نشان داد که رویکرد افرازی و روشK-means همچنان الویت اول خوشه بندی می باشد.

واژه‌های کلیدی: ارزیابی رویکردها و روشها، خوشه بندی، تاپسیس، تست تئوری.

[1] A_Sarafrazi@pnu.ac.ir

۱- مقدمه

امروزه خوشه‌بندی به عنوان یک روش یادگیری بدون ناظر در کاربردهای بسیاری توانسته است ارزش خود را نشان دهد. در این مجموعه سعی شده تا حد امکان مطالب پایه‌ایِ خوشه‌بندی و مسائل مربوط به آن بیان شود. همچنین سعی شده است تا چندین روش و تکنیک مختلف و رایجِ خوشه بندی تشریح شود و ویژگی‌های هر یک بیان گردد. برای ارزیابی، سنجش و اعتبارسنجی خوشه‌های تولید شده که خود یکی از مسائل مهم و قابل گسترش در باب خوشه‌بندی است، نیز مطالبی گردآوری شده است که امید است مورد توجه خوانندگان قرار گیرد. یکی از روشهای حیاتی کنترل و مدیریت داده ها، کلاس بندی یا گروه بندی داده های با خواص مشابه درون مجموعه ای از دسته ها یا خوشه ها می باشد. مسئله اصلی، گروه بندی بدون نظارت مجموعه ای از الگوهای بدون برچسب، درون گروه های معنی دار است. پس خوشه بندی منجر به فشرده سازی و کاهش اطلاعات می شود(گالو،۲۰۱۱، میلا، ۲۰۱۱، جین،۲۰۰۹ و مورتی۱۹۹۹).

برای این منظور لازم است که الگوهای با بیشترین میزان شباهت، در یک خوشه قرار گیرند. معیارهای مختلفی برای اندازگیری شباهت مورد استفاده قرار می گیرند که وابسته به ماهیت داده ها هست. K- means محبوب ترین روش خوشه بندی موجود الگوریتم می باشد(جین، ۲۰۰۹). این الگوریتم خوشه ای کریسپ و کروی شکل را بر اساس کمینه سازی مجموع مجذور مربعات خطا بین الگوها تا مرکز نزدیکترین خوشه پیدا می کند. پس شکل خوشه ها نمی تواند با پراکندگی و نوع داده ها تنظیم شود. همچنین اگر چند داده در فاصله تقریبا مساوی از مرکز چند خوشه قرار داشته باشند، خوشه بندی به درستی صورت نمی پذیرد. برای حل این مشکلات الگوریتم های ارائه شدند که توانایی پیدا نمودن خوشه های مختلفی مانند فازیFCM با اشکال مختلف، داده ها(جین، ۲۰۰۹، سپهر ۱۳۸۶). همچنین دیگر مشتقات و توسعه هایی برای این الگوریتمها بوجود آمد که قادر بودند برای انواع داده مثل داده های سمبلیک و یا داده ها با مقادیر جا افتاده مشکل اساسی این, خوشه بندی مناسبی را محقق سازند. این قبیل الگوریتم ها، وابستگی دقت آنها به تعداد و مکان اولیه مراکز خوشه ها است. برخی از الگوریتم های تکاملی توانستند این قبیل مشکلات را با ادغام و انشعاب خوشه ها و جابجایی مراکز خوشه ها در حین تولید نسل های الگوریتم تکاملی، حل نمایند(جین، ۲۰۰۹، سپهر ۱۳۸۶). این الگوریتم ها با کدگذاری به شکل رشته ای از بیتها، فضای داده ها را تقسیم بندی می کنند.

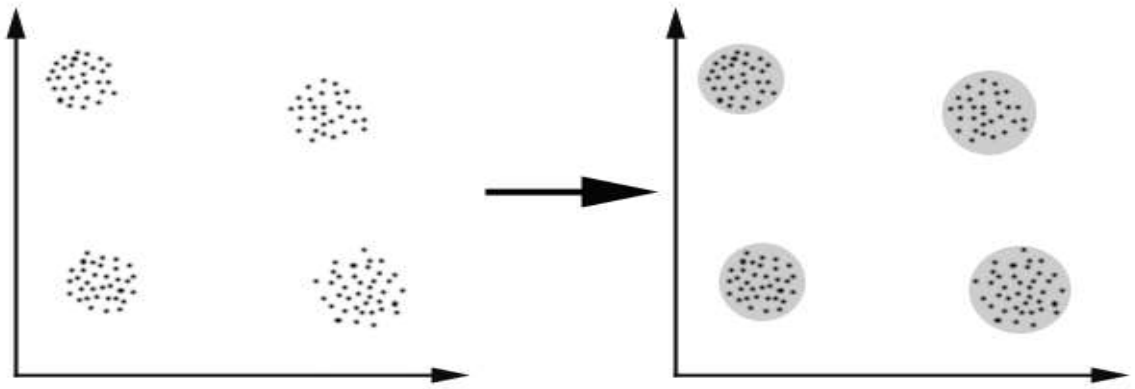
پیشرفت های سریع علمی و تکنولوژی های ذخیره سازی داده ها و رشد دراماتیک در موضوعاتی همانند جستجوهای اینترنتی تصویر سازی دیجیتالی, ویدوئی و تحقیقات بیولوژیکی و ژنتیکی مجموعه داده هایی با حجم و ابعاد بسیار بالا تولید می کند. پیش بینی شد که در سال۲۰۰۷ حجم مصرف داده های دیجیتالی جهانی در حدود ۲۸۱ اگزا[2]بایت و در سال ۲۰۱۱ در حدود ۱۰برابر سال ۲۰۰۷ می شود(جین، ۲۰۰۹، جانتز ۲۰۰۸). بررسی ها نشان داد که بیشترین حجم داده ها توسط رسانه های دیجیتالی تولید و ذخیره می شوند بنابراین بایستی روشهای مناسبی برای تحلیل خودکار, طبقه بندی و بازیابی داده ها ارائه شود. به علاوه رشد زیاد و حجیم داده ها و تنوع آنها همانند متن, تصویر, تصاویر ویدئویی رو به افزایش است. دوربین های ارزان قیمت حجم زیادی از تصاویر و فیلمهای دیجیتالی تولید و آرشیو می کنند. ایمیل ها, بلاگ ها, تراکنش ها, میلیاردها صفحات وب هر روز تترا ها بایت داده را تولید می کند. بسیاری از این داده ها جریانی، غیر ساختار یافته اند و تحلیل آنها را بسیار دشوار می کند. با افزایش حجم و تنوع داده ها نیاز به ایجاد و توسعه متدولوژی هایی است که داده ها را بفهمد و طبقه بندی و خلاصه نماید ضروری است. تکنیک های تجزیه و تجلیل داده ها به طور وسیعی به دو دسته اصلی تقسیم بندی می شوند(جین، ۲۰۰۹، مورتی ۱۹۹۹، توکی ۱۹۷۷).

[2] Exabyte: 1 Exabyte=10^16 or 1000,000 terabytes

در میان منابعی که سازمان‌ها در اختیار دارند، محققان زیادی به اهمیت ویژه منابع انسانی در سازمان‌ها اشاره کرده‌اند. باکسِال و همکاران چنین عنوان کرده است که مهمترین دارایی هر سازمان صبح زود وارد سازمان می‌شود و در انتهای زمان کاری از آن خارج می‌شود (باکسِل و دیگران، ۲۰۰۷). به گفته بوکسال و پارسل (۲۰۰۳)، پفر (۱۹۹۸)، و گراتون و همکاران (۲۰۰۰)، کارکنان سازمان‌ها و مدیریت آن‌ها عنصری اساسی در دست‌یابی به مزیت رقابتی برای سازمان‌ها می‌باشد(باکسِل و پورسِل، ۲۰۰۳، فِفِر، ۱۹۹۸، گِراتون و دیگران، ۲۰۰۰). با توجه به رقابت روز افزون سازمان‌ها و تغییرات محیطی، سازمان‌ها در جست‌وجوی روش‌هایی هستند که بتوان با کمک آن‌ها از طریق منابع انسانی خود به مزیت رقابتی دست یابند. در این میان می‌توان به غنی‌سازی شغل اشاره کرد. غنی‌سازی شغل منجر به فراهم کردن مسئولیت‌ها و چالش‌های شغلی بیشتری برای کارکنان می‌شود. غنی‌سازی شغل این امکان را برای کارکنان مهیا می‌کند تا در کارهایشان اختیارات لازم را برای تصمیم‌گیری داشته باشند. غنی‌سازی شغل به‌طور مستقیم با عوامل انگیزشی و نیز رضایت کارکنان مرتبط می‌باشد. تحقیقات حاکی از آن است که غنی‌سازی شغل منجر به افزایش رضایت شغلی، و انگیزش کارکنان، و نیز کاهش غیبت کاری کارکنان می‌شود (ارپِن، ۱۹۷۹).

۲. خوشه بندی[3]

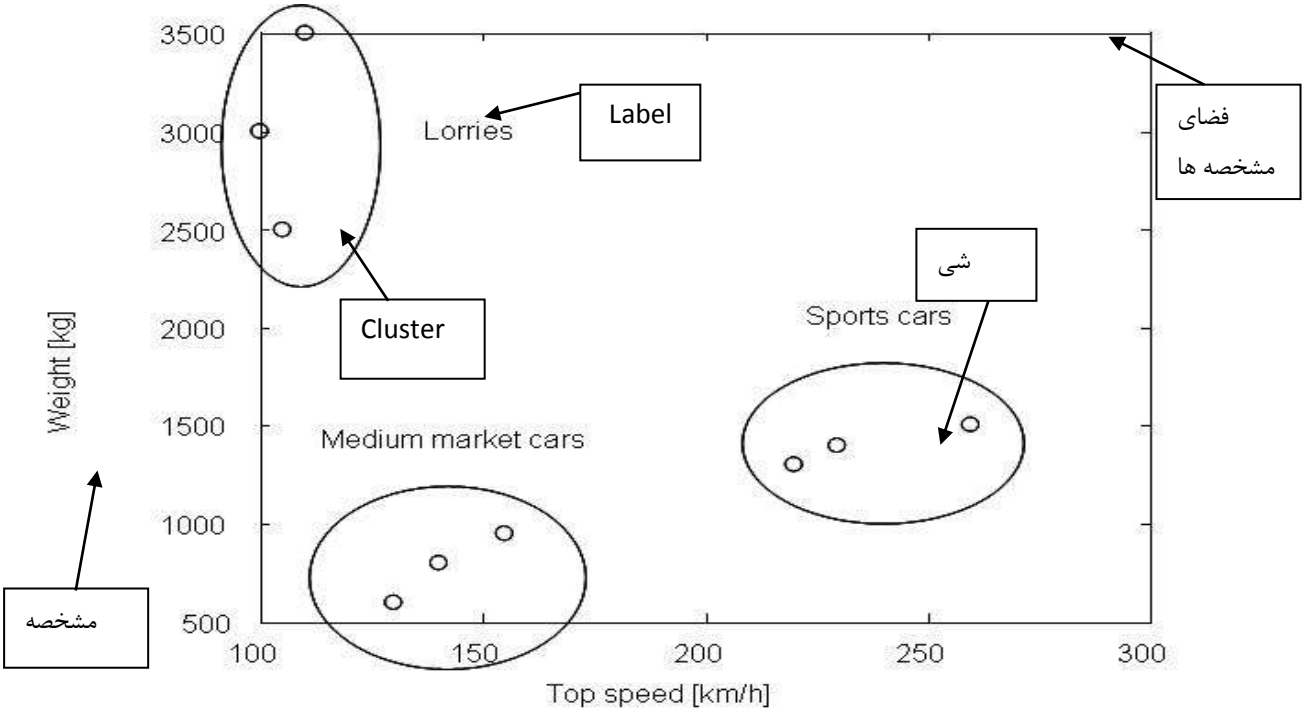
خوشه بندی یکی از شاخه های یادگیری بدون نظارت می باشد و فرآیند خودکاری است که در طی آن، نمونه ها به دسته هایی که اعضای آن مشابه یکدیگر می باشند تقسیم می شوند که به این دسته ها خوشه گفته می‌شود. هدف از خوشه بندی و تجزیه و تحلیل خوشه ای گروه بندی طبیعی از مجموعه الگوها, اشیاء و نقاط است بنابراین خوشه مجموعه ای از اشیاء می باشد که در آن اشیاء با یکدیگر مشابه بوده و با اشیاء موجود در خوشه های دیگر غیر مشابه می باشند..برای مشابه بودن می توان معیارهای مختلفی را در نظر گرفت مثلا می توان معیار فاصله را برای خوشه بندی مورد استفاده قرار داد و اشیائی را که به یکدیگر نزدیکتر هستند را بعنوان یک خوشه در نظر گرفت که به این نوع خوشه بندی، خوشه بندی مبتنی بر فاصله[4]نیز گفته می شود. بعنوان مثال در شکل(۱)نمونه های ورودی در سمت چپ به چهار خوشه مشابه شکل سمت راست تقسیم می شوند. در این مثال هر یک از نمونه های ورودی به یکی از خوشه ها تعلق دارد و نمونه ای وجود ندارد که متعلق به بیش از یک خوشه باشد (جین، ۲۰۰۹، مورتی، ۱۹۹۹).



شکل ۱- خوشه بندی نمونه های ورودی در این شکل نمونه‌ای از اعمال خوشه‌بندی روی یک مجموعه از داده‌ها مشخص شده است که از معیار فاصله به عنوان عدم شباهت بین داده‌ها استفاده شده است (مورتی، ۱۹۹۹).

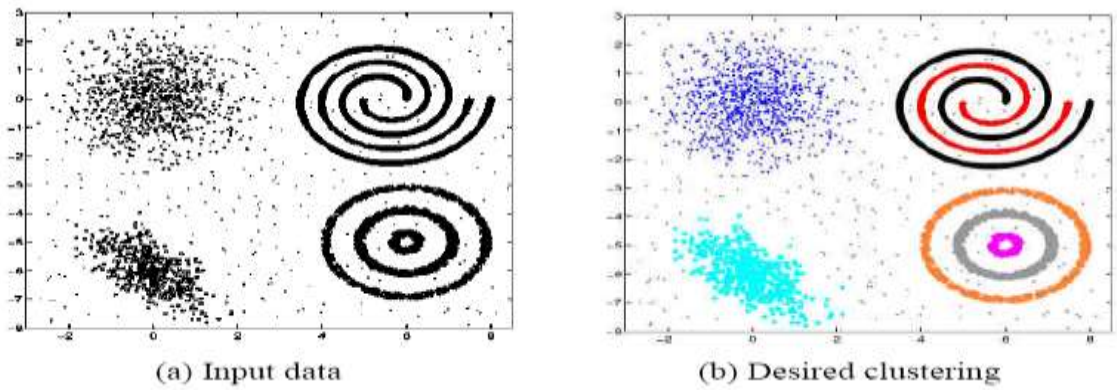
[3] Clustering
[4] Distance-based Clustering

نمونه دیگری در شکل(۲) در نشان داده شده است که در این شکل هر یک از دایره های کوچک یک وسیله نقلیه (شیء) را نشان می دهد که با ویژگی های وزن و حداکثر سرعت مشخص شده اند. هر یک از بیضی ها یک خوشه می باشد عبارت کنار هر بیضی برچسب آن خوشه را نشان می دهد.



شکل ۲- خوشه بندی وسایل نقلیه (مورتی،۱۹۹۹)

همانطور که در شکل دیده می شود وسایل نقلیه به سه خوشه تقسیم شده اند برای هر یک از این خوشه ها می توان یک نماینده در نظر گرفت مثلا می توان میانگین وسایل نقلیه باری را محاسبه کرد و بعنوان نماینده خوشه وسایل نقلیه باری معرفی نمود. در واقع الگوریتم های خوشه بندی اغلب بدین گونه اند که یک سری نماینده اولیه برای نمونه های ورودی در نظر گرفته می شود و سپس از روی میزان تشابه نمونه ها با این نماینده های مشخص می شود که نمونه به کدام خوشه تعلق دارد بعد از این مرحله نماینده های جدید برای هر خوشه محاسبه می شود و دوباره نمونه ها با این نماینده ها مقایسه می شوند تا مشخص شود که به کدام خوشه تعلق دارند این کار آنقدر تکرار می شود تا زمانیکه نماینده های خوشه ها تغییری نکنند. با توجه به شکل(۳) زیر داده های بدون برچسب را نشان می دهد..



شکل ۳- تنوعی از خوشه ها از لحاظ شکل, اندازه و چگالی(جین، ۲۰۰۹)

شکل- الف در شکل- ب نیز هدف توسعه الگوریتمی خودکار برای کشف گروه بندی طبیعی می باشد. در شکل (۳) زیر تنوعی از خوشه ها را نشان می دهد. هفت خوشه موجود در شکل a که با رنگهای متفاوت در شکل b نشان داده شده است از لحاظ شکل, اندازه و چگالی متفاوتند. اگرچه که این خوشه ها در تحلیل داده ها آشکار هستند اما روش و الگوریتم جامعی وجود ندارد که تمامی این خوشه ها را بتواند پیدا کند (جین، ۲۰۰۹). خوشه بندی با طبقه بندی[5]متفاوت است. در طبقه بندی نمونه های ورودی برچسب گذاری شده اند ولی در خوشه بندی نمونه های ورودی دارای بر چسب اولیه نمی باشند در واقع با استفاده از روشهای خوشه بندی است که داده های مشابه مشخص و بطور ضمنی برچسب گذاری می شوند. می توان قبل از عملیات طبقه بندی داده ها یک خوشه بندی روی نمونه ها انجام داد و سپس مراکز خوشه های حاصل را محاسبه کرد و یک بر چسب به مراکز خوشه ها نسبت داد و سپس عملیات طبقه بندی را برای نمونه های ورودی جدید انجام داد.

۲-۲- رویکردها و روش های خوشه بندی

رویکردهای اصلی خوشه بندی عبارتند از: افرازی، سلسله مراتبی، مبتنی بر چگالی، مبتنی بر مشبک کردن فضا، نقشه های خود سازمانده، متاهیورستیک.

۲-۲-۱-روشهای افرازی

افراز از این داده های اشیا درست می کند به طوریکه هر افراز یک خوشه شی داریم یک روش افرازی،n فرض کنید یک پایگاه داده با k گروه خوشه بندی شده و دارای دو شرط زیر می باشد: . پس داده های اشیا در K=<nرا نشان می دهد:

- هرگروه حداقل یک شیء دارد.
- هر شیء تنها به یک گروه تعلق دارد. (این شرط در روشهای افرازی فازی می تواند قابل انعطاف باشد)

فرض کنیم یک پایگاه داده بانn شی داریم.علاوه بر آن تعداد خوشه هایی که باید تشکیل شوند نیز معلوم است . یک الگوریتم افرازی ،اشیاء را در k افراز سازماندهی کرده به طوریکه هر افراز یک خوشه را نمایش می دهد.خوشه ها معمولا با معیاری که تابع شباهت نیز نام دارند شکل میگیرد.بنابراین اشیاءداخل یک خوشه به هم شبیه اند ودر مقابل اشیاء در خوشه های مختلف به هم شبیه نیستند.

این شباهت و عدم شباهت اشیاء بر مبنای داده های پایگاه داده تعیین می شوند. دو الگوریتم مهم این روش عبارتند از: k-means و K-medoids

۲-۲-۲-روشهای سلسله مراتبی

این روش ساختاری سلسله مراتبی از اشیاء یک مجموعه معلوم ایجاد می کند. روش سلسله مراتبی می تواند خوشه بندی را به صورت تجمعی و یا به صورت تقسیمی انجام دهد. به رویکرد تجمعی رویکرد پایین به بالا نیز گفته می شود. این روش با شکل دهی گروهای مجزا که هر یک شامل حداقل یک شیء می باشند شروع می شود. سپس اشیا یا گروه های نزدیک به هم رایکی می کند تا این که در نهایت یک گروه کلی در بالاترین سطح ایجاد شود. در روش تقسیمی کل اشیاء در یک خوشه در نظر گرفته شده و در هر تکرار یک خوشه به دو خوشه ی کوچکتر تقسیم می شود.

۲-۲-۳-روش مبتنی بر چگالی

بسیاری از روشهای افرازی اشیا را براساس فاصله آنها نسبت به یکدیگر خوشه بندی میکنند. برخی روشها تنها خوشه های کرو شکل را پیدا میکنندو در برابر خوشه هایی به شکل های دلخواه با مشکل مواجه میشوند. در مقابل برخی روشهای دیگر خوشه بندی بر پایه ی چگالی توسعه یافتند. ایده عمومی این روشها رشد دادن خوشه ها بر پایه چگالی در همسایگی آنهاست. به این

[5] Classification

معنی که برای هر نقطه داده در یک خوشه معلوم همسایه ای با شعاع مشخص در نظر گرفته می شود. این نوع خوشه بندیبرای هموار سازی اغتشاشات و کشف خوشه هایی با اشکال دلخواه به کار می رود . برخی الگوریتم های مبتنی بر چگالی عبارتند از:

همانطور که در روشهای قبل به خصوص روشهای افرازی مشاهده شد، خوشه های حاصل از این روش ها اغلب دارای شکل هایی متقارن در فضای مسئله بودند.بدین صورت که اغلب حول یک مرکزیت (مثلا میانگین متغیرهای یک خوشه و یا عنصری که به شکل دایره ای ، کروی و... را تشکیل می دادند.گاه ممکن medoid عنوان مرکزیت خوشه آن خوشه انتخاب شده بود یعنی است بنا به ماهیت مسئله به دنبال خوشه هایی با الگوهایی پیچیده تر باشیم و یا اینکه رابطه خاص بین ابعاد مختلف داده ها و متغیرها وجود داشته باشد و به دنبال یافتن عناصری باشد که چنین خصوصیتی را دارد. در این حالت از روش میتنی بر چگالی استفاده می کند. ایده اصلی این روش ها بر این اساس است که ابتدا به دنبال نقاطی می گردد که چگالی حول آنها زیاد باشد سپس سعی می کند به گونه ای نقاطی را که با این مراکز تجمع در ارتباط هستند ،پیدا کند گاه پس از طی چند مرحله ۲ یا چند مرکز تجمع به یکدیگر متصل شده و یک خوشه را شکل می دهند. این روشها هم چنین در حذف داده های پرت و مغشوش بسیار مفید هستند(کینوئن، ۲۰۱۱).

۴-۲-۲- روشهای مبتنی بر مشبک کردن فضا

روش مشبک سازی فضا به سلولهای مختلف امکان کار بر روی اطلاعات با درجه تفکیک شفافیت های متفاوت را فراهم میکند در این روش ابتدا فضا به سلول هایی تقسیم شده و سپس عملیات خوشه بندی روی این سلول ها انجام می گیرد. مهم ترین مزیت این روش افزایش سرعت است. زیرا پیچیدگی محاسباتی را کاهش می دهد چرا که پیچیدگی وابسته به تعداد سلول هاست نه تعداد داده ها است. در این روش فضا به سلول هایی STING ابتدایی ترین و ساده ترین روش در این دسته روش شبکه اطلاعات آماری یا ابتدایی تقسیم می شود.اغلب اوقات از روی این سلول ها سلول هایی دیگر در لایه ای بالاتر تشکیل می شوندیعنی مثلا ازترکیب هر۴سلول، یک سلول در لایه ای بالاتر با درجه تفکیک کمتر شکل می گیرد و این کار به صورت سلسله مراتبی برای چندین لایه تکرار می شود.

سپس برای هر سلول اطلاعات آماری مانند میانگین ، میانه ، بیشینه، کمینه ،انحراف معیار استاندارد و... محاسبه می شود. این پارامترهای آماری و حتی نوع توزیع آماری داده های پایگاه داده محاسبه شده و به هر سلول تخصیص داده می شوند. چنین توزیع می تواند توسط کاربر مشخص شده و یا توسط امتحان فرضیه هایی مانند تستX^2 معین شوند کاملا مشخص است که اطلاعات مرتبه های بالاتر از مرتبه های پایین تر به سادگی قابل محاسبه خواهند بود.

۵-۲-۲-نقشه های خود سازمانده

ابزار قدرتمند و جذابی برای نمایش داده های چند بعدی در فضاهای با ابعاد پایین، (SOM) نقشه های خود سازمان یا سازمانده روشی برای خوشه بندی و پیش پرداز اطلاعات می باشد. نقشه های SOM (معمولا یک یا دو بعد)فراهم می کندو همچنین خود سازمانده که گاهی نقشه های مشخصه خود سازمان و یا نقشه های کوهونن نامیده می شود،توسط پروفسور تیووکوهونن از دانشگاه فنلاند ابداع شده است. این فرایند کاهش بعد بردارها ،روشی برای فشرده سازی داده ها به نام کمی سازی برداری شبکه ای برای ذخیره اطلاعات ایجاد می کند به نحوی که ارتباط مکانی بین مجموعه آموزشی حفظSOM می باشد. علاوه بر این، با شبکه رقابتی عبارت است از:SOM می شود. تفاوت هیچ سوگیری وجود ندارد سوگیری مقدار وزن نرون ورودی ثابت است. SOM در علاوه بر نرون برنده ، نرو نهای همسایه نیز تطبیق یافته و اوزان آن ها اصلاح می شود. ،نگاشت رنگ ها در صفحه دو بعدی است .فرض کنید هزاران مشاهده داریم و SOM مثالی متداول برای کمک به آموزشی مبانی هر مشاهد یکی از ۸ رنگ سمت راست می باشد. میگوتی، ۲۰۰۶ مطرح کرد که هر رنگ از سه جز قرمز، سبز وآبی تشکیل شده است که می توانند دارای مقادیری بین ۰ تا ۲۵۵ باشند ،بنابراین هر مشاهده دارای سه ویژگی می باشند.

۶-۲-۲-رویکرد خوشه بندی فازی[۶]

در واقع در اینجا با مفهوم عدم قطعیت[۷]مواجه هست. بشر نیز از عدم قطعیت در زندگی روزمره بارها استفاده کرده است مثلا هوای سرد، آب داغ و غیره. تفاوت اصلی مجموعه های فازی و مجموعه های کلاسیک در این است که تابع تعلق مجموعه های فازی دو مقداری نیست (۰ یا ۱) بلکه می تواند هر مقداری بین ۰ تا ۱ را اختیار کند. حال مجموعه انسانهای جوان و پیر را در نظر بگیرد اگر ۲۵ سال را سن جوانی در نظر بگیرد می تواند به ۲۵ تعلق ۱ بدهد و مثلا به ۳۰ تعلق ۰,۸ و به ۳۵ تعلق ۰,۷۵ و به ۹۰ تعلق ۰,۱ را بدهد. اگر اعضای یک مجموعه فازی تنها دارای تابع تعلق ۰ و ۱ باشند این مجموعه فازی یک مجموعه کلاسیک خواهد بود(ناسیمنتو، ۲۰۰۶، بزدک،۱۹۹۸، غضفری،۱۳۸۷).

در خوشه بندی کلاسیک هر نمونه ورودی متعلق به یک و فقط یک خوشه می باشد و نمی تواند عضو دو خوشه و یا بیشتر باشد. مثلا هر یک وسایل نقلیه عضو یک خوشه می باشد و نمونه ای عضو دو خوشه نیست و به زبان دیگر خوشه ها همپوشانی ندارند. حال حالتی را در نظر بگیرید که میزان تشابه یک نمونه با دو خوشه و یا بیشتر یکسان باشد در خوشه بندی کلاسیک باید تصمیم گیری شود که این نمونه متعلق به کدام خوشه است. تفاوت اصلی خوشه بندی کلاسیک و خوشه بندی فازی در این است که یک نمونه می تواند متعلق به بیش از یک خوشه باشد.

۷-۲-۲-رویکرد خوشه بندی متاهیورستیک[۸]

برای مسائلی با ابعاد بزرگ، روش سیمپلکس از کارایی بسیار خوبی برخوردار است، ولی روش شاخه و کرانه کارایی خود را از دست می‌دهد و عملکرد بهتری از شمارش کامل نخواهد داشت. به دلایل فوق، اخیراً تمرکز بیشتری بر روش‌های ابتکاری[۹]یا فرا ابتکاری یا جستجوی تصادفی[۱۰]صورت گرفته است. روش‌های جستجوی ابتکاری، روش‌هایی هستند که می‌توانند جوابی خوب (نزدیک به بهینه) در زمانی محدود برای یک مسأله ارائه کنند. روش‌های جستجوی ابتکاری عمدتاً بر مبنای روش‌های شمارشی می‌باشند، با این تفاوت که از اطلاعات اضافی برای هدایت جستجو استفاده می‌کنند. این روش‌ها از نظر حوزه کاربرد، کاملاً عمومی هستند و می‌توانند مسائل خیلی پیچیده را حل کنند. عمده این روش‌ها، تصادفی بوده و از طبیعت الهام گرفته شده‌اند (رضا عالی تپه، یقینی،۱۳۸۶).

مسائل بهینه‌سازی ترکیبی[۱۱]در طول دو دهه گذشته، کاربرد بهینه‌سازی در زمینه‌های مختلفی چون مهندسی صنایع، برق، کامپیوتر، ارتباطات و حمل و نقل گسترش یافته است. بهینه‌سازی خطی و غیرخطی (جستجو جهت یافتن مقدار بهینه تابعی از متغیرهای پیوسته)، در دهه پنجاه و شصت از اصلی‌ترین جنبه‌های توجه به بهینه‌سازی بود. بهینه‌سازی ترکیبی عبارت است از جستجو برای یافتن نقطه توابع با متغیرهای گسسته و در دهه ۷۰ نتایج مهمی در این زمینه به دست آمد. امروزه بسیاری از مسائل بهینه‌سازی ترکیبی (مانند مسأله فروشنده دوره‌گرد) که اغلب از جمله مسائل سخت NP-hard هستند، به صورت تقریبی (نه به طور دقیق) در کامپیوترهای موجود قابل حل می‌باشند. در این الگوریتم‌ها هیچ ضمانتی برای آنکه جواب به دست آمده بهینه باشد، وجود ندارد و تنها با صرف زمان بسیار می‌توان جواب نسبتاً دقیقی به دست آورد؛ در حقیقت با توجه به زمان صرف شده، دقت جواب تغییر می‌کند. برای روش‌های ابتکاری نمی‌توان تعریفی جامع ارائه کرد. با وجود این، در اینجا کوشش می‌شود تعریفی تا حد امکان مناسب برای آن عنوان شود: روش جستجوی ابتکاری، روشی است که می‌تواند جوابی خوب (نزدیک به بهینه) در زمانی محدود برای یک مسأله ارائه کند. هیچ تضمینی برای بهینه بودن جواب وجود ندارد و متأسفانه نمی‌توان میزان نزدیکی جواب به دست آمده به جواب بهینه را تعیین کرد.در سال‌های اخیر یکی از مهمترین و امیدبخش‌ترین تحقیقات، روش‌های ابتکاری برگرفته از طبیعت بوده است؛ این روش‌ها شباهت‌هایی با سیستم‌های اجتماعی و

---

[6] Fuzzy
[7] Uncertainty
[8] Heuristic-Meta
[9] Heuristic
[1] Random Method [0]
[1] Optimization Problems Combinational

یا طبیعی دارند. کاربرد آنها برگرفته از روش‌های ابتکاری پیوسته می‌باشد که در حل مسائل مشکل ترکیبی نتایج بسیار خوبی داشته است (ابراهیمی،۱۳۸۶، طاهریان فرد، ۱۳۸۷).

۳-۲- معیارهای ارزیابی

معیارهای ارزیابی مورد نظر شامل موارد ذیل است. البته معیارهای دیگری نیز می تواند مد نظر قرار گیر که در اینجا به ۱۹ معیار با اهمیت آن اشاره شده است (علی احمدی، ۱۳۸۶):

۱- معیار منحصر به فرد بودن: بدین معنی که یک تئوری باید منحصر به فرد و یگانه بوده و متمایز از تئوری دیگری باشد.

۲- معیار محا فظه کارانه: یک تئوری جدید نمی تواند جایگزین تئوری موجود گردد مگر اینکه تئوری جدید مشخصا دارای دلایل قانع کننده و خصوصیات برتری نسبت به تئوری موجود باشد.

۳- معیار قا بلیت تعمیم: تعداد حوزه هایی که در آ« می تواند صادق بوده و کاربرد داشته باشد، بسیار مهم است لذا چنانچه یک تئوری موضوعات بیشتری را نسبت به تئوری دیگری تحت پوشش قرار دهد از اهمیت بیشتری برخوردار است.

۴- معیار قا بلیت تولید ایده: چنانچه یک تئوری قابلیت تولید مدلها و فرضیات بیشتری را داشته باشد ارجح تر خواهد بود.

۵- معیار محدودیت تئوری: چنانچه تئوری فرضیات محدو کننده کمتری داشته باشد از برترتی بیشتری نسبت به تئوری که فرضیات محدود کننده متعددی دارد بر خوردار است به عبارت دیگر تئوری باید عاری از اجزاء مازاد باشد.

۶- معیار سادگی و اثربخشی: سادگی بیان و درک تئوری ویژگی است که یک تئوری را نسبت به تئوریهای پیچیده و دشوار در انتقال مفهوم متمایز می سازد به عبارت دیگر تئوری خوب باید با حداقل روابط و متغیرهای کلیدی، پدیده و خروجی مورد نظر را توصیف نماید.

۷- معیار سازگاری و ثبات داخلی: بدین معنی است که متغیرها، اجزاء و عواملی که تئوری در بردارد با یکدیگر به صورت منطقی هماهنگ و سازگار باشد.

۸- معیار ریسک اجرا و تجربه: آزمونها و تجربیات مبتنی بر تئوری عمدتا با مخاطره مواجه هستند. لذا چنانچه تئوری به گونه ای توسعه یابد که ریسک و مخاطره اجرا را کاهش دهد مطلوبتر خواهد بود به عبارت دیگر تئوری باید ابعاد اجرایی و تجربی را نیز در برگیرد.

۹- معیار مجرد از زمان و مکان: بدین معنی که تئوری باید مستقل از زمان و مکان بوده و همواره بر اساس روابط و متغیرهای آن صادق باشد.

۱۰- معیار شفاف و بدون ابهام: بدین معنی که تئوری باید کاملا واضح و منظور و مقصد آن آشکار باشد به نحوی که امکان تعریف و ارائه تعابیر و تفاسیر مختلف فراهم نشود.

۱۱- معیار مطلوبیت خروجی و نتایج: خروجی و نتایج تئوری باید به موضوعات جالب، زیبا و مهیج منتهی گردد نه اینکه بیان دیگری باشد از موضوعاتی که قبلا به آنها پرداخته شده است.

۱۲- معیار قدرت توجیه گری:کوششی برای توضیح وقایع گذشته است فرضیه های علی انتظاراتی را درباره روابط میان متغیرها ایجاد می نماید می توانیم چنین انتظاراتی را به عنوان قدرت توجیه گری در نظر گرفت.

۱۳- معیار برنامه پیشرو در برابر انحطاط: تمایز بین برنامه های پژوهشی پیشرو و در حال انحطاط به طور وسیعی مربوط به شمول و دامنه محدودیتهایی است که از طریق فرضهای کمکی بر تئوری اعمال شود.

۱۴- معیارکاربرد: باید تئوری را که علی رغم ثابت بودن عوامل دیگر کاربردهای متعددی را نتیجه می دهد به آنکه کاربردهای کمتری را نتیجه می دهد ترجیح داد.

۱۵- معیار تحت تاثیر قرار دادن اندیشمندان: نظریه جدید بر اندیشه غالب افراد مطلع و متخصص تاثیری شگرف بر جای گذارد و آنان را به واکنش عملی وا دارد.

۱۶- معیار صرفه جویی: در صرفه جویی تئوریها باید ساده سازی شوند و فرضهای غیر ضروری باید برچیده شوند.

۱۷- معیار توصیف همه جانبه: یک بینش همه جانبه می تواند هم توصیفی باشد و هم هنجاری.

۱۸- معیار واقهیت گرایی: هدف بنیادین در تحقیقات این است که واقعیت توضیح داده شود شاید از این روست که برخی می گویند، بدون نظریه فهم واقعیت غیر ممکن است.

۱۹- معیارنقش مدلها در القای نظریه: نظریه ها بدون مدلها، قادر به تبیین پدیده ها نیستند و مدلها نیز با قیاس و تمثیل این کار را انجام می دهند. در این تحقیق تعدادی از مهمترین معیارها انتخاب گردید و در بررسی رویکردها و روشهای تئوری خوشه بندی لحاظ شد.

۳- روش شناسی تحقیق

۳-۱- روش تحقیق

تحقیق حاضر از نظر هدف، کاربردی است و از نظر روش گردآوری داده ها توصیفی است.

۳-۲- ابزار جمع آوری اطلاعات

بدلیل بهره مندی کامل از داده و اطلاعات در تحقیق حاضر از ترکیب روشهای جمع آوری داده ها استفاده شده است. ادبیات موضوع با روش کتابخانه ای و اینترنتی و گردآوری داده ها جهت تجزیه و تحلیل با پرسشنامه تکمیل شده توسط ۳نفر از خبرگان این حوزه، انجام شد.

۳-۳- جامعه آماری

در نیم قرن اخیر روش های مختلف بررسی الگوی دادهها و کشف دانش بر اساس تجزیه و تحلیل چند متغیره توسعه روشهای بسیاری از خوشه بندی را توسعه داده است. بنابراین جامعه آماری مد نظر برای مطالعه منتخب از ۷ رویکرد و ۲۱ روش خوشه بندی داده ها ست که از الگوریتم های معروف و پر کاربرد در پیشینه تحقیق انتخاب شد.

۳-۴- روش تجزیه و تحلیل

روش تاپسیس[1]TOPSIS به علت وجود چهار مزیت ذیل استفاده شده است:

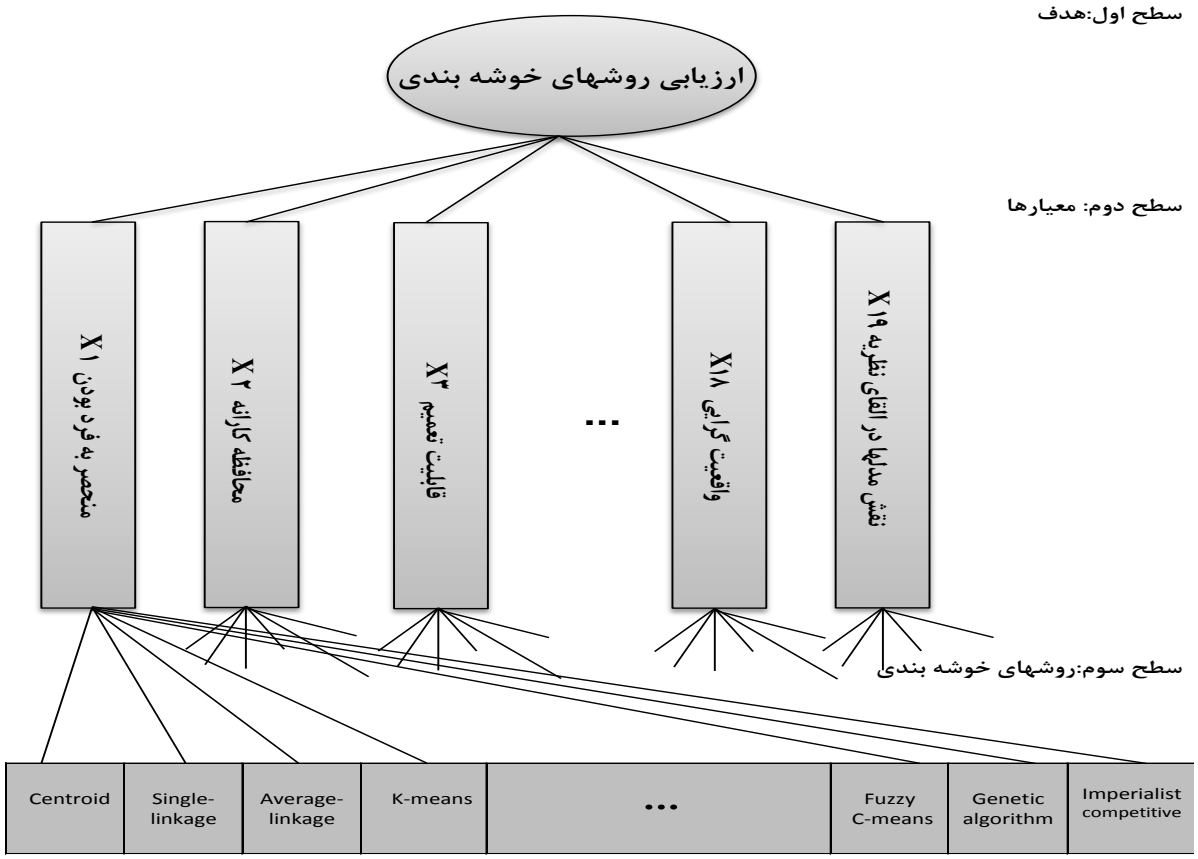
۱-دارا بودن استدلالی معتبر که به خوبی منطق انتخاب افراد را تشریح می کند

۲- محاسبه ارزش عددی برای بهترین و بدترین آلترناتیوها

۳-دارا بودن فرآیند محاسباتی ساده ای که به راحتی در صفحات گسترده قابل برنامه نویسی است

[1] Technique for Order Preference by Similarity to the Ideal solution

۴-عملکرد چندوجهی آلترناتیوها در معیارها(حداقل در دو وجه قابل تصور است) شده است، در میان هشت روش مقایسه ای که توسط زاناکیس و همکارانش دار ای کمترین نقص در رتبه بندی TOPSIS گروه مدل های جبرانی ارزیابی چندمعیاره روش آلترناتیوها می باشد. با توجه به مرور ادبیات موضوع ۱۶ متغیر معیار تعریف شد که تلاش شد تا پوشش قابل قبولی از ویژگیهای یک خوشه فناوری را در بر داشته باشد. خوشه های صنعتی فعال نیز به عنوان گزینه های تصمیم گیری و الویت بندی و تحلیل نزدیکی به وضعیت آرمانی درنظر گرفته شد(شکل-۵).



شکل ۴- قالب پژوهشی ارزیابی روشهای خوشه بندی

۵-۳- تحلیل با استفاده از تصمیم گیری های چند معیاره

تصمیم‌گیری‌ها بر دو دسته هستند که دسته اول تصمیم‌گیری بر اساس چند معیار و دسته دوم تصمیم‌گیری بر اساس چند هدف متفاوت است. MCDM معمولاً برای انتخاب بهترین گزینه ارائه شده استفاده می‌شود که ممکن است معیارهای آنها با یکدیگر در تعارض باشد. MCDM که تصمیم‌گیری چند هدفه است، می‌تواند به طور همزمان بر چند که هدف متناقض هستند تمرکز کرده با روش‌های برنامه‌ریزی ریاضی بهترین راه حل را ارائه دهد. (زنجیرانی، ۱۳۸۵) MCDM به برتری نسبی اهداف و ارتباط بین اهداف و شاخص‌ها توجه می‌کند. (یانگ، ۲۰۰۷). MCDM برای انتقال بهترین گزینه از بین گزینه‌های پیشنهاد شده با توجه به شاخص‌های ارزیابی هر گزینه به کار می‌رود (زنجیرانی) MCDM به دلیل داشتن معیارهای ذهنی یک رویکرد توصیفی است. هدف MCDM تعیین بهترین گزینه در حالی که بتواند بیشترین رضایتمندی را ایجاد کند (یانگ و همکارانش،۲۰۰۷). روش‌های ترکیبی و روش‌های فاصله‌ای و روش‌های برتری نسبی از جمله روش‌های رایج MCDM است (رومرو،۲۰۰۰). بلتون و همکارانش یک دسته بندی اما در دهه های اخیر توجه محققین معطوف به مدلهای تصمیم گیری

چند معیاره برای تصمیم گیر ی های پیچیده شده است .در این تصمیم گیری ها به جای استفاده از یک معیار سنجش از چندین معیار سنجش استفاده به عمل آید (اصغرپور، ۱۳۷۷).

۶-۳- تکنیک TOPSIS

در تصمیم گیریهای چندمعیاره به جای استفاده از یک معیار سنجش بهینگی از چندین معیار سنجش ممکن است استفاده گردد این مدلهای تصمیم گیری به دو دسته کلی مدلهای چند هدفه و مدلهای چند شاخصه تقسیم می شوند در این پروژه که از مدلهای چند شاخصه استفاده گردیده به طور کلی جهت انتخاب مناسبترین گزینه از بین m گزینه موجود می باشد لذا تصمیم گیریهای چند شاخصه بصورت ماتریس نمایش داده می شود که تعداد سطرهای مبین گزینه های موجود و تعداد ستونها مبین شاخصهای موجود می باشد. جهت اولویت بندی گزینه های موجود از روش TOPSIS استفاده گردیده که الگوریتم روش به شرح ذیل می باشد: در این روش با در نظر گرفتن فاصله یک گزینه A_i از نقطه ایده آل، فاصله آن از نقطه ایده آل منفی هم در نظر گرفته می شود. بدان معنی که گزینه انتخابی باید دارای کمترین فاصله از راه حل ایده آل بوده و در عین حال دارای دورتین فاصله از راه حل ایده آل منفی باشد. در این روش مطلوبیت هر شاخصی باید به طور یکنواخت افزایش یا کاهش (یعنی هرچه r_{ij} بیشتر؛ مطلوبیت بیشتر و یا برعکس) باشد. الگوریتم روش به قرار زیر می باشد:

گام ۱:تبدیل ماتریس تصمیم گیری موجود به یک ماتریس بی مقیاس شده

گام ۲: ایجاد ماتریس «بی مقیاس وزین» با مفروش بودن بردار w به عنوان ورودی به الگوریتم

$W= \{w_1, w_2, \ldots w_n\} \cong$ (مفروض از DM)

ماتریس بی مقیاس وزین $V = N_D \times W_{n\times n}$

N_D : ماتریسی است که امتیازات شاخصها در آن بی مقیاس شده و قابل مقایسه شده است $W_{n\times n}$ ماتریسی است قطری که فقط عناصر قطراصلی آن غیر صفر است.

گام ۳: مشخص نمودن گزینه ایده آل (A^+) و ایده آل منفی

(A^-) گزینه ایده آل=

$A^+ =\{(\max V_{ij} | j\in j),(\min V_{ij} | J\in j`) | i=1,2,\ldots m\}$

$=\{V^+_1, V_2^+, \ldots, V^+i, \ldots, V^+_n\}$

=گزینه ایده آل منفی

$A^- =\{(\min V_{ij} | j\in j),(\max V_{ij} | J\in j`) | i=1,\ldots,m\}=$

$=\{V^-_1, V^-_2, \ldots V^-_n\}$

J={j=1,2,…,n| jها مربوط به سود باشند}

j`={j=1,2,…n,| jها مربوط به هزینه}

گام ۴:محاسبه اندازه جدایی: فاصله گزینه iام با ایده آل با استفاده از روش اقلیدسی بدست می آید.

d^+_i=فاصله گزینه iام از ایده آل $= \left\{\sum_{j=1}^{n}\left(v_{ij} - v^+ j\right)^2\right\}^{0.5}$

$i = 1,2,\ldots m$

d^-_i=فاصله گزینه iام از ایده آل منفی $= \left\{\sum_{j=1}^{n}\left(v_{ij} - v - j\right)^2\right\}^{0.5}$

$i = 1,2,\ldots m$

گام ۵: محاسبه نزدیکی نسبی A_1 به راه حل ایده آل

$$cli = \frac{d_1^-}{d_1^- + d_1^+} \quad 0 \le cli \le 1 \qquad i = 1,2,...m$$

گام ۶: رتبه بندی گزینه ها: براساس ترتیب نزولی cli^+ می توان گزینه های موجود از مسأله مفروض را رتبه بندی کرد. لذا بزرگترین مقدار cli^+ بهترین گزینه خواهدبود (اصغرپور، ۱۳۷۷).

۴-تجزیه و تحلیل

۴-۱- تجزیه و تحلیل رویکردهای خوشه بندی

با توجه به نظر سنجی از خبرگان, متوسط نظرات خبرگان در جدول ۱ آورده شد. جدول تصمیم ذیل ماتریسی با ابعاد ۱۹*۷ در خصوص بررسی رویکردهای در خوشه بندی است. سطر های ماتریس شامل انواع رویکرهای مختلف به تحلیل خوشه ای داده ها می باشد و ستونهای ماتریس نیز شامل ۱۹ معیاری است که بر اساس آن معیارها مناسبترین رویکرد و روش تجزیه و تحلیل از منظر تست تئوری بررسی می شود. معیارهای و متغیرهای موجود در جدول زیر در ستونها آمده است(علی احمدی، ۱۳۸۶). با بررسی های بیشتر ممکن است معیارهای بیشتریرا توسعه داد که در اینجا تلاش شد مهمترین آنها در مدل وارد شود. در روش جمع آوری داده ها به هر معیار وزنی بر اساس مقیاس ۱-۱۰۰ داده شد و همچنین با مقیاس ۱-۹ اهمیت هر معیار را برای هر رویکرد یا روش تخصیص یافت.

جدول۱- ماتریس تصمیم رویکردهای خوشه بندی[۱-۴۲]

رد	معیار / تئوری/رویکرد	محصر به فرد بودن	محافظه کارانه	قابلیت تعمیم	قابلیت تولید ایده	محدودیت تئوری	سادگی و اثربخشی	سازگاری و ثبات داخلی	ریسک اجرا و تجربه	محرز از زمان و مکان	شفاف و بدون ابهام	مطلوبیت خروجی و نتایج	قدرت توجیه گری	برنامه پیشرو در برابر انحطاط	کاربردها	تحت تاثیر قرار دادن اندیشمندان	صرفه جویی	توصیف همه جانبه	واقعیت گرایی	نقش مدلها در القای نظریه
		x1	x2	x3	x4	x5	x6	x7	x8	x9	X 10	X 11	X 12	X 13	X 14	X 15	X 16	X 17	X 18	X 19
	وزن	۸۵	۳۵	۸۵	۹۵	۶۰	۸۵	۷۰	۶۵	۹۵	۷۵	۹۰	۷۰	۷۵	۸۰	۹۵	۵۰	۶۵	۴۵	۷۰
۱	رویکرد سلسله مراتبی	۵	۷	۷	۳	۵	۵	۷	۵	۷	۵	۵	۷	۵	۹	۵	۷	۵	۷	۵
۲	رویکرد افرازی	۷	۵	۹	۹	۷	۷	۷	۷	۹	۷	۷	۹	۹	۹	۹	۷	۹	۷	۹
۳	رویکرد فازی	۷	۵	۷	۵	۷	۳	۷	۷	۷	۵	۹	۷	۷	۷	۷	۵	۷	۹	۹
۴	رویکرد مبتنی بر چگالی	۷	۳	۷	۳	۵	۳	۵	۵	۷	۵	۵	۷	۵	۷	۵	۵	۵	۷	۷
۵	رویکرد مشبک کردن فضا	۷	۳	۵	۳	۵	۱	۵	۵	۷	۳	۷	۷	۵	۵	۵	۳	۵	۷	۵
۶	رویکرد خود سازمانده	۵	۱	۵	۳	۳	۳	۷	۵	۷	۳	۷	۷	۷	۵	۵	۳	۵	۷	۷
۷	رویکرد متاهیورستیک	۹	۵	۹	۹	۷	۵	۵	۹	۷	۵	۷	۹	۹	۹	۹	۵	۵	۷	۹

سپس جدول تصمیم را بر اساس الگوریتم تاپسیس نرمالیزه و موزون نموده و نتایج جدول تصمیم نرمال شده جهت تهیه جدول گزینه های آرمانی استفاده می شود. در این بخش هر معیار بر اساس ماهیت آن در تحلیل و با توجه به موضوع تحلیل

مقدار حداکثر یا حداقل را دریافت می کند. جدول ۲ گزینه های آرمانی مثبت و منفی را برای رویکرهای متعدد خوشه بندی به نمایش گذاشته است. بخش اول جدول شامل معیارهای آرمانی مثبت می باشد که به غیر از متغیر ۵ و ۸ یعنی محدویت تئوری و ریسک اجرا سایر معیارها مثبت بودن آنها در جهت حداکثر سازی و دو معیار مذکور مثبت بودن آن در جهت منفی است. در ادامه در بخش دوم جدول گزینه های آرمانی منفی در جهت مخالف قبل آمده است و در سلول های ماتریس آنها مقادیر حداقل و حداکثری داده ها قرار می گیرد.

جدول ۲- گزینه های آرمانی جدول تصمیم

A+	Max	Max	Max	Max	Min	Max	Max	Min	Max	Max	Max	Max	Max	Max	Max	Max	Max	Max	Max	Max
A+	0.03037	0.01463	0.02752	0.04099	0.00849	0.03789	0.02126	0.01407	0.03161	0.02926	0.03236	0.02231	0.02656	0.0264	0.03469	0.01823	0.0223	0.01488	0.02276	
A-	Min	Min	Min	Min	Max	Min	Min	Max	Min	Min	Min	Min	Min	Min	Min	Min	Min	Min	Min	Min
A-	0.01687	0.00209	0.01529	0.01366	0.0198	0.00541	0.01519	0.02533	0.02459	0.01254	0.01798	0.01735	0.01475	0.01467	0.01927	0.00781	0.01239	0.01157	0.01265	

سپس بر اساس رابطه $i=1,2,...m$ $0 \le cci \le 1$ $cci = \frac{d_1^-}{d_1^- + d_1^+}$ ضریب نزدیکی مثبت هر گزینه تصمیم در خصوص رویکردهای خوشه بندی در جدول ۳ خلاصه شد.

جدول ۳- جدول ضریب نزدیکی و فاصله رویکردهای خوشه بندی تا شرایط آرمانی

گزینه های خوشه بندی	ضریب نزدیکی		رویکرد
A1	0.43	cc1+	رویکرد سلسله مراتبی
A2	0.77	cc2+	رویکرد افرازی
A3	0.46	cc3+	رویکرد فازی
A4	0.32	cc4+	رویکرد مبتنی بر چگالی
A5	0.54	cc5+	رویکرد مشبک کردن فضا
A6	0.30	cc6+	رویکرد خود سازمانده
A7	0.64	cc7+	رویکرد متاهیورستیک

۲-۴- تجزیه و تحلیل روشها و الگوریتمهای خوشه بندی

جدول تصمیم تجزیه و تحلیل روشهای خوشه بندی همانند تحلیل رویکردهای آن دارای ماتریسی با ابعاد ۱۹*۲۱ در خصوص بررسی رویکردهای در خوشه بندی است. سطر های ماتریس شامل ۲۱ الگوریتم منتخب و پر کاربرد نسبت به سایر الگوریتمها بوده است و ستونهای ماتریس نیز شامل ۱۹ معیاری است که بر اساس آن معیارها مناسبترین رویکرد و روش تجزیه و تحلیل از منظر تست تئوری بررسی می شود. معیارهای و متغیرهای موجود در جدول ۴ در ستونها آمده است.

در ادامه بر اساس روش ارائه شده در مقاله و همانند بخش اول با پردازش بر روی داده های جدول تصمیم جدول۵ و ۶ را که شامل گزینه های آرمانی تصمیم می باشد تعیین و میزان آمال حداکثری و حداقلی را برای گزینه های آرمانی مثبت و منفی و رویکرهای متعدد خوشه بندی به نمایش گذاشته است. بخش اول جدول شامل معیارهای آرمانی مثبت می باشد که به غیر از

متغیر ۵ و ۸ یعنی محدویت تئوری و ریسک اجرا سایر معیارها مثبت بودن آنها در جهت حداکثر سازی و دو معیار مذکور مثبت بودن آن در جهت منفی است. جدول ۶ نیز مقادیر ضریب نزدیکی گزینه های روشهای خوشه بندی را نمایش می دهد.

جدول۴ - ماتریس تصمیم روشها و الگوریتم های خوشه بندی[۱-۴۲]

رد	معیار / تئوری/متدلوژی/الگوریتم	محصر به فرد بودن	محا فظه کارانه	قا بلیت تعمیم	قا بلیت تولید ایده	محدودیت تئوری	سادگی و اثربخشی	سازگاری و ثبات داخلی	ریسک اجرا و تجربه	مجرد از زمان و مکان	شفاف و بدون ابهام	مطلوبیت خروجی و نتایج	قدرت توجیه گری	برنامه پیشرو در برابر انحطاط	کاربردها	تحت تاثیر قرار دادن اندیشمندان	صرفه جویی	توصیف همه جانبه	واقعیت گرایی	نقش مدلها در القای نظریه
		x1	x2	x3	x4	x5	x6	x7	x8	x9	x 10	x 11	x 12	x 13	x 14	x 15	x 16	x 17	x 18	x 19
	وزن	۸۵	۳۵	۸۵	۹۵	۶۰	۸۵	۷۰	۶۵	۹۵	۷۵	۹۰	۷۰	۷۵	۸۰	۹۵	۵۰	۶۵	۴۵	۷۰
۱	Centroid	۵	۵	۳	۳	۱	۳	۵	۳	۵	۵	۳	۳	۱	۳	۳	۵	۱	۳	۱
۲	Single-linkage	۵	۷	۷	۳	۳	۵	۵	۵	۵	۷	۵	۵	۳	۳	۳	۵	۳	۵	۵
۳	Average-linkage	۷	۷	۷	۳	۳	۵	۵	۵	۵	۷	۵	۵	۳	۳	۳	۵	۳	۵	۵
۴	Complete-linkage	۷	۷	۷	۵	۳	۵	۵	۵	۵	۷	۵	۵	۳	۳	۳	۵	۳	۵	۵
۵	Wards	۷	۷	۷	۵	۵	۵	۷	۷	۵	۷	۷	۷	۵	۵	۵	۵	۳	۵	۵
۶	K-means	۹	۹	۷	۷	۹	۷	۹	۹	۹	۷	۷	۹	۹	۹	۹	۵	۷	۷	۹
۷	K-median	۷	۵	۷	۵	۵	۳	۵	۳	۳	۵	۷	۳	۵	۵	۵	۳	۳	۳	۵
۸	X-means	۷	۵	۵	۵	۵	۵	۵	۵	۵	۳	۷	۵	۵	۳	۵	۳	۳	۵	۵
۹	G-means	۷	۷	۵	۵	۵	۵	۵	۵	۵	۳	۷	۵	۵	۳	۵	۳	۳	۵	۵
۱۰	GX-means	۷	۷	۷	۵	۳	۵	۵	۵	۵	۳	۷	۵	۵	۳	۳	۳	۳	۵	۵
۱۱	OPTIC	۷	۵	۷	۳	۵	۳	۵	۳	۵	۳	۷	۷	۵	۳	۳	۳	۵	۵	۵
۱۲	Fuzzy C-means	۹	۷	۹	۷	۷	۵	۷	۹	۹	۵	۹	۷	۷	۷	۷	۵	۷	۷	۷
۱۳	Density –base SCAN	۷	۵	۵	۵	۵	۵	۵	۵	۷	۳	۷	۵	۷	۵	۵	۵	۳	۵	۵
۱۴	Grid-base	۷	۵	۷	۵	۳	۳	۵	۵	۷	۳	۷	۵	۵	۵	۳	۳	۱	۵	۵
۱۵	Self-Organizing map	۷	۷	۵	۵	۵	۵	۷	۷	۷	۳	۷	۷	۷	۵	۵	۵	۵	۵	۷
۱۶	GA: Genetic algorithm	۹	۷	۹	۷	۷	۷	۷	۷	۹	۵	۹	۷	۹	۹	۹	۵	۷	۷	۷
۱۷	SA: Simulant Annealing	۹	۵	۷	۵	۵	۵	۵	۵	۷	۵	۹	۷	۵	۵	۵	۳	۵	۵	۵
۱۸	PSO: Partial Swarm Optima	۹	۵	۷	۵	۷	۷	۷	۵	۷	۵	۹	۷	۷	۵	۷	۳	۵	۵	۷
۱۹	TS: Tabu Search	۷	۵	۷	۵	۳	۵	۵	۵	۷	۵	۷	۷	۵	۵	۳	۳	۵	۵	۵
۲۰	ACO: Ant Colony Optima	۷	۵	۵	۵	۳	۵	۵	۵	۵	۵	۷	۷	۵	۵	۳	۳	۵	۵	۵
۲۱	ICA: Imperialist competitive	۹	۵	۵	۵	۵	۵	۷	۷	۵	۵	۹	۷	۵	۵	۵	۳	۵	۵	۷

جدول ۵- گزینه های آرمانی جدول تصمیم

A+		Max	Max	Max	Max	Min	Max	Max	Min	Max	Max	Max	Max	Max	Max	Max	Max	Max	Max	Max	Max
A+		0.017023	0.00797	0.017894	0.020615	0.017018	0.018493	0.01671	0.016164	0.021361	0.016371	0.017824	0.01602	0.018903	0.022607	0.026311	0.009636	0.016352	0.011789	0.017294	
A-		Min	Min	Min	Min	Max	Min	Min	Max	Min	Min	Min	Min	Min	Min	Min	Min	Min	Min	Min	Min
A-		0.009457	0.004428	0.005965	0.008835	0.001891	0.007926	0.005283	0.005388	0.00712	0.007016	0.005941	0.00534	0.0021	0.007536	0.00877	0.005782	0.002336	0.00393	0.001922	

جدول ۶- جدول ضریب نزدیکی و فاصله رویکردهای خوشه بندی تا شرایط آرمانی

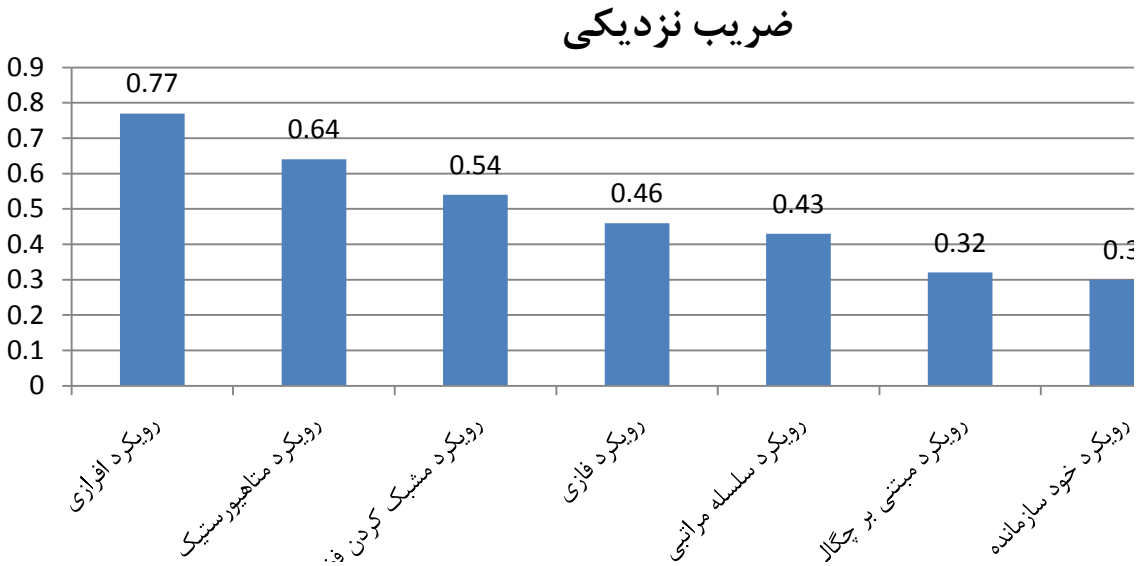
متدلوژی/روش		ضریب نزدیکی	گزینه های خوشه بندی
Centroid	cc1+	0.350	A1
Single-linkage	cc2+	0.410	A2
Average-linkage	cc3+	0.413	A3
Complete-linkage	cc4+	0.422	A4
Wards	cc5+	0.496	A5
K-means	cc6+	0.853	A6
K-median	cc7+	0.431	A7
X-means	cc8+	0.443	A8
G-means	cc9+	0.444	A9
GX-means	cc10+	0.425	A10
OPTIC	cc11+	0.417	A11
Fuzzy C-means	cc12+	0.670	A12
Density –base SCAN	cc13+	0.475	A13
Grid-base	cc14+	0.424	A14
Self-Organizing map	cc15+	0.512	A15
GA: Genetic algorithm	cc16+	0.734	A16
SA: Simulant Annealing	cc17+	0.502	A17
PSO: Partial Swarm Optima	cc18+	0.570	A18
TS: Tabu Search	cc19+	0.462	A19
ACO: Ant Colony Optima	cc20+	0.449	A20
ICA: Imperialist competitive	cc21+	0.504	A21

۵- نتیجه گیری و تحقیقات آتی

خوشه بندی یکی از پرکاربرد ترین روشهای علمی تحلیل است که تقریبا در اکتشافات علمی در اکثر حوزه های و رشته های علمی مورد استفاده قرار می گیرد در این مقاله تلاش شد که رویکردها و روشهای پرکاربرد خوشه بندی مورد تحلیل و ارزیابی قرار گیرد. رویکردهای مورد بررسی به ترتیب الویت از منظر شاخص های مورد نظر به ترتیب شامل رویکرد افرازی، متاهیورستیک، مشبک کردن فضا ،فازی ،سلسله مراتبی ،مبتنی بر چگالی ،خود سازمانده می باشد که البته می توان رویکردهای دیگری نیز جستجو نمود اما بیشترین رویکردها، مورد توجه قرار گرفت. جدول ۷ و نمودار ۱ نتایج را به ترتیب الویت نمایش می دهد.

جدول ۷- رتبه بندی رویکردهای خوشه بندی

	نام رویکرد	الویت بندی	گزینه خوشه بندی	ضریب نزدیکی /فاصله
۱	رویکرد افرازی	a1	A2	0.77
۲	رویکرد متاهیورستیک	a2	A7	0.64
۳	رویکرد مشبک کردن فضا	a3	A5	0.54
۴	رویکرد فازی	a4	A3	0.46
۵	رویکرد سلسله مراتبی	a5	A1	0.43
۶	رویکرد مبتنی بر چگالی	a6	A4	0.32
۷	رویکرد خود سازمانده	a7	A6	0.30



نمودار ۱- رتبه بندی رویکردهای خوشه بندی

همچنین با رتبه بندی روشهای مختلف که زیر مجموعه هایی از رویکرها می باشند در جدول ۸ و نمودار ۲ نمایش داده شد.

جدول ۸- رتبه بندی رویکردهای خوشه بندی

	روش ها و الگوریتم های خوشه بندی	الویت بندی	گزینه های خوشه بندی	ضریب نزدیکی /فاصله
1	K-means	a1	A6	0.853
2	GA: Genetic algorithm	a2	A16	0.743
3	Fuzzy C-means	a3	A12	0.670
4	PSO: Partial Swarm Optima	a4	A18	0.570
5	Self-Organizing map	a5	A15	0.512
6	ICA: Imperialist competitive	a6	A21	0.504
7	SA: Simulant Annealing	a7	A17	0.502
8	Wards	a8	A5	0.496
9	Density –base SCAN	a9	A13	0.475
10	TS: Tabu Search	a10	A19	0.462
11	ACO: Ant Colony Optima	a11	A20	0.449
12	G-means	a12	A9	0.444
13	X-means	a13	A8	0.443
14	K-median	a14	A7	0.431
15	GX-means	a15	A10	0.425
16	Grid-base	a16	A14	0.424
17	Complete-linkage	a17	A4	0.422
18	Average-linkage	a18	A3	0.413
19	OPTIC	a19	A11	0.417
20	Single-linkage	a20	A2	0.410
21	Centroid	a21	A1	0.350

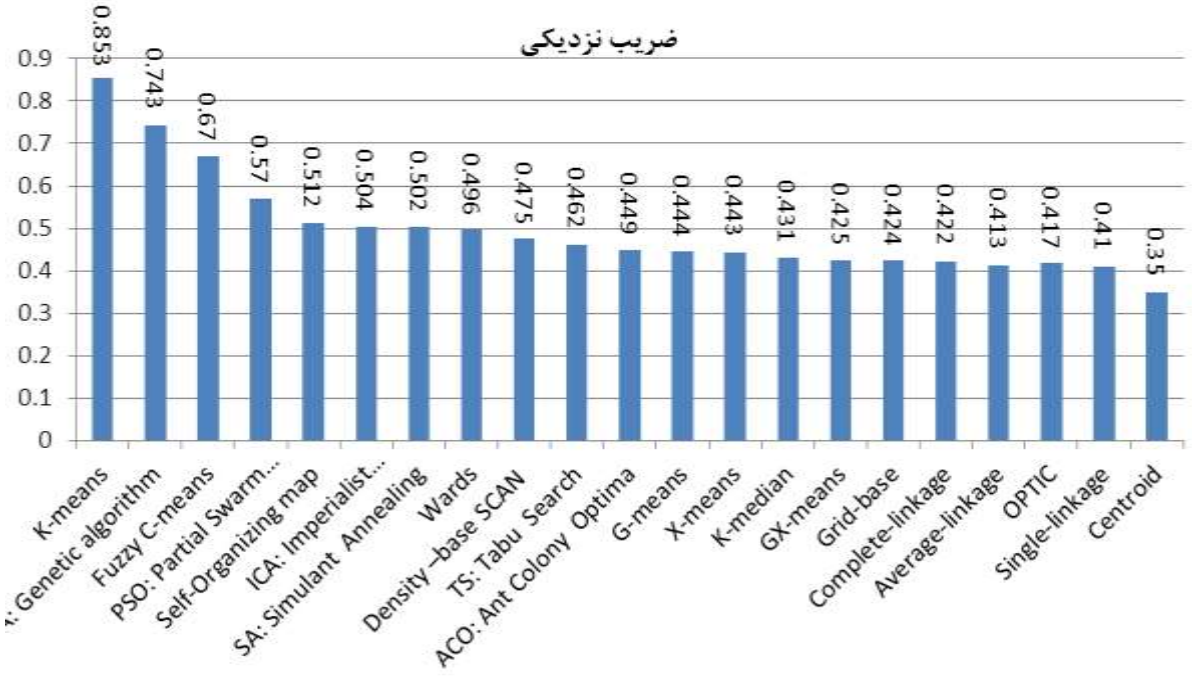
بر اساس نتایج بدست آمده روش K-means با بیش از ۵۰ سال توسعه در رتبه اول قرار دارد که بیشترین توسعه و رایج ترین و مشهورترین روش می باشد. سپس روشهای مبتنی بر الگوریتم ژنتیک و فازی به ترتیب در رتبه دوم و سوم قرار گرفته اند. قابل توجه که بسیاری از این روشها در دهد اخیر با ترکیب دو روش بهبود یافته اند و پژوهشگران تلاش کردند که نقاط ضعف هر روش را بهبود داده و از نقاط قوت در الگوریتم های ترکیبی بهره گیرند. رویکرد ترکیبی موجب توسعه روشهای بسیاری در حل مسئله خوشه بندی گردید که می تواند در تحقیقات آتی روشهای ترکیبی را تحلیل یا با رویکردهای دیگر مقایسه نمود.

در هر حال به دلیل کاربرد بسیار وسیع خوشه بندی در توسعه علمی و کشف دانش روشهای آن به شکلی پویا در حال توسعه است. اما در مجموع همه روشهای فوق سعی در کاهش پیچیدگی مسائل و کشف روابط متغیرها مشابه در خصوص تصمیم گیری در برابر مسائل مختلف توسعه یافته اند همچنین این روشها در برابر مسائل متعدد با محدویتها و مشخصات مختلف عملکردهای متعددی از خود نشان می دهند. با ملاحضه جدول۹ نتایج پژوهشی در این مقاله مقایسه شده و توسط هامودا و کارای ۲۰۱۰ و جین نیل ۲۰۰۹ تایید می شود.

جدول ۹- نتایج مقایسات تجربی چهار روش و الگوریتم (هامودا، ۲۰۱۰)

الگوریتم	جنبه های مورد مقایسه			
	RMSE	Accuracy%	Regression Line slope	Time sec
K-means	0.447	80.0	0.600	0.9
Fuzzy C-means	0.469	78.0	0.559	2.2
Mountain	0.469	78.0	0.556	118
subtractive	0.500	75.0	0.507	3.6

صحت مقایسه نتایج از روش های تصمیم گیری تاپسیس و پارامترهای عملکرد چهار الگوریتم انجام شده است اما در این مطالعه ۲۱ الگوریتم پرکاربرد از بین الگوریتم های مختلف و توسعه یافته روشهای خوشه بندی در نیم قرن اخیر انتخاب و بررسی شد. خلاصه نتایج در نمودار ۲ به صورت نزولی نشان داده شده است. در هر حال خوشه بندی از تئوریها و روشهایی است که دائما در حال توسعه است و کمک های زیادی به اکتشاف حقایق و دانش در انبوهی از انواع داده تولید شده امروزی خواهد نمود که الگوریتم های مختلف آن در صدد کاهش هزینه و بهبود عملکرد آن هستند.



نمودار ۲-نتایج رتبه بندی روشهای خوشه بندی

۶- منابع

[1]Gullo, Francesco, Domeniconi, Carlotta, Tagarelli, Andrea," Advancing Data Clustering via Projective Clustering Ensembles", SIGMOD'11, June 12–16, 2011, Athens, Greece
[2]Meila, Marina, Classic and Modern Data Clustering, University of Washington, 2011
[3]JAIN, A.K., MURTY, M.N., FLYNN, P.J., Data Clustering: A Review, ACM Computing Surveys, Vol. 31, No. 3, September 1999
[4]Jain, Anil K., Data Clustering: 50 Years beyond K-Means, to appear in Pattern Recognition Letters, 2009
[5]Westendorf1, Sascha, Literature Review on knowledge
Engineering, Data Clustering and Computational Creativity, Knowledge Engineering, Software Architecture and Design Patterns,
[6]Berkhin, Pavel, Survey of Clustering Data Mining Techniques, Accrue Software, Inc.
[7]Algergawy, Alsayed, Mesiti, Marco, Nayak, Richi, XML Data Clustering: An Overview, ACM Computing Surveys, Vol., No. , 2009
[8] Qiu, Dingxi, Tamhane, Ajit C., A comparative study of the K-means algorithm and the normal
Mixture model for clustering: Univar ate case, Journal of Statistical Planning and Inference 137 (2007) 3722 – 3740
[9] Larocque, Aaron, Valova, Iren, Evaluation of classification quality and comparative analysis of clustering and self-organization, Procedia Computer Science 6 (2011) 141–146
[10]Burke, Edmund, Kendall, Graham, Comparison of Meta-Heuristic Algorithms for Clustering Rectangles,
[11]Kinnunen, Tomi, Sidoroff, Ilja, Tuononen, Marko, Franti, Pasi, Comparison of clustering methods: A case study of text-independent speaker modeling, Pattern Recognition Letters 32 (2011) 1604–1617
[12]Mingoti, Sueli A., Lima, Joab O., Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms, European Journal of Operational Research 174 (2006) 1742–1759
[13]Luxburg, Ulrike von, Ben-David, Shai, Towards a Statistical Theory of Clustering, School of Computer Science, University of Waterloo, Canada
[14]K.Arunprabha, M.C.A.,M.Phil, V.Bhuvaneswari, M.Sc., M.Phil, Comparing K-value Estimation for Categorical and Numeric Data Clustering, International Journal of Computer Applications (0975– 8887),Volume 11– No.3, December 2010
[15]Solka, Jeffrey L., Text Data Mining: Theory and Methods, Statistics Surveys, Vol. 2 (2008) 94–112
[16]ZHANG, TIAN, RAMAKRISHNAN, RAGHU, LIVNY, MIRON, BIRCH: A New Data Clustering lgorithm and Its Applications, Data Mining and Knowledge Discovery, 1, 141–182 (1997)
[17]Alitappeh, Reza, Ebadzadeh, Mohammad Mehdi, Data Clustering Using A New CGA (Chaotic-Generic Algorithm) Approach,
[18]Buan, Aslak, Bakke, Marsh, Robert, Cluster-tilting theory, 2000 Mathematics Subject Classification. Primary 16G20, 16G70, 16S99, 17B37; Secondary 17B20, 52B11.
[19]YANG, M.-S., Survey of Fuzzy Clustering, 1993, Mathl. Comput. Modelling Vol. 18, No. 11, pp. 1-16
[20]Hammouda, Khaled, Karray, Fakhreddine, A Comparative Study of Data ClusteringTechniques,2010, University of Waterloo, Ontario, Canada
[21] Mehdizadeh, Esmaeil, A fuzzy clustering PSO algorithm for supplier base management, 2009, International Journal of Management Science and Engineering Management Vol. 4, No. 4, pp. 311-320
[22]Bataineh, K. M., Naji, M., Saqer, M., A Comparison Study between Various Fuzzy Clustering Algorithms,2011, Jordan Journal of Mechanical and Industrial Engineering, Volume 5, Number 4, Pages 335 – 343
[23]Akthar, Shaheda, Rafi, Sk.Md, IMPROVING THE SOFTWARE ARCHITECTURE THROUGH FUZZYCLUSTERING TECHNIQUE., Indian Journal of Computer Science and Engineering Vol 1 No 1 54-57
[24] Baraldi, Andrea, Blonda, Palma, A Survey of Fuzzy Clustering Algorithms for Pattern Recognition—Part I,1999, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 29, NO. 6
[25]Nascimento, S., Mirkin, B., Moura-Pires, F., Fuzzy Clustering Model of Data and Fuzzy c-Means,
[26]BEZDEK, JAMES C., EHRLICH, ROBERT, FULL, WILLIAM, FCM: THE FUZZY c-MEANS CLUSTERING ALGORITHM, 1984, *Computers & Geosciences* Vol. 10, No. 2-3, pp. 191-203
[27] Park, Han-Saem, Yoo, Si-Ho, Cho, Sung-Bae, Evolutionary Fuzzy Clustering Algorithm with Knowledge-Based Evaluation and Applications for Gene Expression Profiling, 2005, Journal of Computational and Theoretical Nanoscience Vol.2, 1–10
[28]HARTIGAN, JOHN A., Clustering Algorithms, 1975, John Wiley & Sons, Inc.
[29]Kaufman, leonard, Rousseeuw, peter J., Finding Groups in Data An introduction to cluster Analysis, 2005, John Wiley & Sons, Inc.

[۳۰]علی احمدی، علیرضا، سعید نهایی، وحید، ۱۳۸۶،توصیفی جامع از روشهای تحقیق، انتشارات تولید دانش
[۳۱]علی احمدی، علیرضا، ۱۳۸۶،روش تحقیق و راهنمای پایاننامه نویسی، انتشارات تولید دانش
[۳۲]اصغرپور، محمد، ۱۳۷۷، تصمیم گیری چند متغیره، انتشارات دانشگاه تهران
[۳۳]یقینی، مسعود، غضنفری، ناهید،ارائه یک الگوریتم ترکیبی برای خوشه بندی با استفاده از رویکرد فرا ابتکاری جستجوی ممنوع
[۳۴]دانش، ملیحه، یغمایی مقدم، محمدحسین،اکبرزاده توتونچی، محمدرضا، ۱۳۸۹، خوشه بندی داده با استفاده از ترکیب،PSO,K-harmonic means، هجدهمین کنفرانس مهندسی برق ایران، اصفهان
[۳۵]حاجی مهدیزاده زرگر، سمانه، بزرگ نیا، آرزو، یغمائی مقدم، محمد حسین، الگوریتم FIGKM رهیافتی برخوشه بندی بهینه
[۳۶]عسگریان، احسان, معین زاده، حسین، سریانی، محسن،حبیبی، جعفر، رویکرد جدید برای خوشه بندی فازی بوسیله الگوریتم ژنتیک
[۳۷]ابراهیمی، افشین، اله کبیر، احسان، کارو لوکس، انتخاب ویژگی با الگوریتم وراثتی برای خوشه بندی قلم های فارسی با FCM
[۳۸]طاهریان فرد، الهه، کارو لوکس، نجار اعرابی، بابک، روشی جدید در خوشه بندی اطلاعات با استفاده ازترکیب الگوریتم کشورهای استعماری و k-means
[۳۹]مختاری حسن آباد، وحید، کنگاوری، محمدرضا، ۱۳۸۶،خوشه بندی داده های جریانی با استفاده از موازی سازی الگوریتم های ترکیبی، دانشگاه آزاد اسلامی قزوین
[۴۰]سپهر، ریحانه، مرادی، محمد حسن، مشایخی، غنچه، ۱۳۸۶، بررسی و مقایسه رو شهای مختلف خوشه بندی فازی تفکیکی مبتنی بر روش استاندارد خوشه بندی فازیFCM هفتمین کنفرانس سیستمهای فازی، دانشگاه فردوسی مشهد
[۴۱]ظهیری, حمید، ۱۳۸۵، کاربرد الگوریتم ایمنی مصنوعی در خوشه یابی داده ها، چهارمین کنفرانس ماشین بینایی و پردازش تصویر ایران، دانشگاه فردوسی مشهد
[۴۲]غضنفری، مهدی، علیزاده، سمیه، تیمور پور، بابک، داده کاوی و کشف دانش، انتشارات دانشگاه علم و صنعت ایران، ۱۳۸۷چاپ اول

فصلنامه پژوهش در علوم، مهندسی و فناوری
دوره ۴، شماره ، ۱۳۹۷، صفحات ۶۵-۸۴

Half a century after clustering; reviewing and evaluating approaches and clustering methods with multi-criteria decision analysis

Abbas Sarafrazi*
A_sarafrazi@pnu.ac.ir
Faculty member, Ph.D Student, Industrial Engineering, Payam Noor Universiry, Tehran, Iran

Abstract
Clustering as an unmonitered learning method in today's many applications has shown its value. One of the vital ways of controlling and managing data, classifying or grouping data with similar properties are within a set of categories or clusters. For this purpose, it is necessary that the patterns with the highest similarity are in a cluster. The main clustering approaches are: partitioning, hierarchical, density-based, space-based, self-organizing maps, and Heuristic-Meta. Multi-criteria models are used in this project. In general, there are several index available to select the most suitable indices. Therefore, multi-criteria decisions are displayed in a matrix, with the number of columns indicating the indices and existing criteria. The TOPSIS method was then used to prioritize the available criteria. According to an expert opinion poll, the average expert opinion in the 7*19 matrix decision table was developed for reviewing approaches and 19*21 for clustering methods. The matrix rows include various types of approaches to the cluster analysis of the data and the matrix columns also include 19 criteria based on which criteria are considered the most appropriate approach and analysis method from the perspective of the theory test. The results showed that the partitioning approach and K-means method are still the first priority of clustering.

Keywords: Evaluation of approaches and methods, Clustering, TOPSIS, Theory test.