



دانشگاه آزاد اسلامی

واحد ساری

پایان نامه برای دریافت درجه کارشناسی ارشد «M.CS»

رشته : علوم کامپیوتر

عنوان :

جستجوی مشتریان وفادار بر اساس الگوریتم خوشه بندی

مبتنی بر ژنتیک بهینه و پیش بینی رفتار مشتریان

استاد راهنما :

دکتر همایون موتمنی

استاد مشاور :

دکتر جواد وحیدی

دانشجو :

عارف یلغی

تابستان ۱۳۹۱

بسم الله الرحمن الرحيم

تابستان ۱۳۹۱



دانشگاه آزاد اسلامی

واحد ساری

ساری

پایان نامه برای دریافت درجه کارشناسی ارشد «M.Sc»

رشته : علوم کامپیوتر

عنوان :

جستجوی مشتریان وفادار بر اساس الگوریتم خوشه بندی مبتنی بر ژنتیک

بهبود و پیش بینی رفتار مشتریان

استاد راهنما :

دکترهما یون موتمنی

استاد مشاور :

دکتر جواد وحیدی

دانشجو :

عارف یلغی

تابستان ۱۳۹۱

«چکیده»

امروزه فرآیند استخراج اطلاعات معتبر، از پیش ناشناخته، قابل فهم و قابل اعتماد از پایگاه داده‌های بزرگ و استفاده از آن در تصمیم‌گیری در فعالیت‌های تجاری بکار برده می‌شود. در تحقیق حاضر ارائه یک روش کارا به منظور خوشه بندی مشتریان با استفاده از خوشه بندی مبتنی بر ژنتیک بهینه خودکار انجام گرفته شده است. و پس از آن با پیشنهاد مفاهیم دسته بندی مشتریان با استفاده از ابزار داده کاوی orange، محصولات مورد علاقه مشتریان مدل سازی شده است.

کلید واژه: خوشه بندی، ژنتیک، کشف قوانین وابستگی، مدیریت روابط مشتری

سپاسگذاری

برای نگارش این پایان نامه از محضر استاد ارجمند جناب آقای دکتر همایون مومنی و دکتر جواد وحیدی بهره بردم که این یادآوری سپاس بی پایان من نسبت به کمک های بی دریغ ایشان است. با این تحقیق و با کمک اساتید ارجمندم افقی تازه ای در علم کامپیوتر برایم بازمانده است که تاثیر کلیدی در مسیر پیشروی من خواهد داشت.

تقدیم

تقدیم به روح بزرگ پدرم و دست های چروکیده و جسم خسته مادرم که بدون آنها هیچوقت در مسیر علم قرار نمی گرفتم.

فرم شماره 5



رشته تحصیلی: علوم کامپیوتر

شماره دانشجویی: ۸۹۰۹۷۴۹۴۶

کد پایان نامه: ۲۰۸۴۱۰۰۶۹۰۱۰۱۰

بسمه تعالی

صورجلسه دفاع از پایان نامه کارشناسی ارشد

با یاری از خداوند متعال جلسه دفاع از پایان نامه کارشناسی ارشد آقای عارف یلمقی

باعنوان: جستجوی مشتری‌بان وفادار بر اساس الگوریتم خوشه بندی مبتنی بر ژنتیک بهینه در تاریخ ۹۱/۴/۸

با حضور هیأت داوری برگزار شد و با درجه عالی با نمره ۱۸.۷۵ به تصویب رسید.

نام و نام خانوادگی

سمت

دکتر همایون مومنی

استاد راهنما

دکتر جواد وحیدی

استاد مشاور

دکتر مجید ابوطالبی

استاد داور

دکتر همایون مومنی

مدیر گروه

دکتر محسن ربانی

معاون پژوهش و فناوری

فصل چاپ با تقدیرش صادر و ثبت شد: ☒ بل ☐ خیر
نمره وقت و ثبت: ۱۷.۲۸... با درجه عالی... تاریخ: ۹۱/۴/۸... اعلام برائت: ☒ بل ☐ خیر
از محل بایگسازی شهر اقدام نمایند



«فهرست مطالب»

صفحه	عنوان
۱	فصل اول : کلیات تحقیق
۲	۱- مقدمه
۳	۲-۱ بیان مساله
۴	۳-۱ اهمیت مساله
۵	۴-۱ اهداف کلی و اختصاصی
۵	۵-۱ سوال های تحقیق
۶	۶-۱ تعاریف تحقیق
۶	۱-۶-۱ مدیریت دانش
۶	۲-۶-۱ داده کاوی
۷	۳-۶-۱ دورنما
۸	۷-۱ مفاهیم خوشه بندی

۸	۱-۷-۱ تعاریف خوشه بندی
۹	۲-۷-۱ شرح مختصر تاریخچه تکنیک خوشه بندی
۱۰	۳-۷-۱ تکنیک خوشه بندی
۱۰	۴-۷-۱ علایم نشانه گذاری
۱۲	۸-۱ جستجوی فرااکشافی ژنتیک
۱۲	۱-۸-۱ مقدمه
۱۴	۲-۸-۱ الگوریتم ژنتیک چیست
۱۷	۳-۸-۱ نمونه کروموزم
۱۹	۹-۱ مدیریت روابط مشتری
۱۹	۱-۹-۱ مقدمه
۱۹	۲-۹-۱ تاریخچه مدیریت روابط مشتری
۲۰	۳-۹-۱ ویژگی های کاربردی CRM
۲۰	۴-۹-۱ ضرورت بکارگیری CRM
۲۱	۵-۹-۱ انواع مشتریان
۲۲	فصل دوم: پیشینه تحقیق
۲۳	۱-۲ روش های خوشه بندی
۲۳	۱-۱-۲ خوشه بندی انحصاری و خوشه بندی هم پوشی
۲۷	۲-۱-۲ روش خوشه بندی k-means
۲۸	۳-۱-۲ مشکلات روش خوشه بندی k-means
۲۸	۴-۱-۲ خوشه بندی با روش وارد
۲۹	۵-۱-۲ خوشه بندی فازی
۳۲	۶-۱-۲ الگوریتم isodata
۳۲	۱-۶-۱-۲ شیوه کار isodata
۳۴	۲-۶-۱-۲ مزایا و معایب
۳۴	۷-۱-۲ الگوریتم isoclus
۳۶	۲-۲ پیشینه جنبه ساختاری مورد مطالعه

۴۱	۳-۲ پیشنهاد جنبه کاربردی مورد مطالعه
۴۱	۱-۳-۲ چهارچوب تئوریک و فرضیات عمده در رابطه با وفاداری مشتریان
۴۳	۲-۳-۲ تحلیل ارزش مشتریان RFM
۴۷	۴-۲ برگرفته از مطالعات
۴۸	فصل سوم: روش انجام کار
۴۹	۱-۳ روش انجام کار
۴۹	۱-۱-۳ مقدمه
۵۱	۲-۱-۳ بررسی گام اول از دید ساختاری
۵۱	۱-۲-۱-۳ الگوریتم خوشه بندی پیشنهادی
۵۲	۲-۲-۱-۳ یافتن تعداد خوشه
۵۳	۳-۲-۱-۳ محاسبه تابع ارزیابی
۵۴	۴-۲-۱-۳ اعمال مراحل ژنتیکی بر روی مجموعه داده و اعتبار سنجی
۶۵	۲-۳ جنبه کاربردی پیشنهادی
۶۵	۱-۲-۳ تعاریف
۶۷	۲-۳-۳ مولفه های بیرونی
۶۹	۳-۳-۳ مولفه های درونی
۷۰	فصل چهارم: مقایسه و نتایج
۷۱	۱-۴ مقدمه
۷۱	۲-۴ مقایسه الگوریتم پیشنهادی از نظر دقت حرکتی
۷۱	۱-۲-۴ نگاهی به جزئیات k-mean
۷۳	۲-۲-۴ نگاهی به جزئیات FCM
۷۴	۳-۲-۴ نگاهی به جزئیات GA پیشنهادی
۷۶	۳-۴ مقایسه مجموع فواصل درستی عملکرد خوشه
۷۷	۴-۴ مقایسه با استفاده از شاخص های خوشه بندی
۷۸	۵-۴ مقایسه نسل ها
۷۹	۶-۴ مهاجرت نسل ها

۸۱	 ۷-۴ مطالعه مورد جنبه کاربردی
۸۸	 ۸-۴ نتیجه کلی از کار پیشنهادی
۸۹	 فصل پنجم: نتیجه گیری و کارهای آتی
۹۰	 ۱-۵ نتیجه گیری کلی
۹۰	 ۲-۵ پیش بینی درباره روندهای آینده نرم افزارهای مدیریت روابط مشتری
۹۳	 منابع

«فهرست جداول»

۱۴	جدول (۱-۱): کاربردهای الگوریتم ژنتیک
۵۵	جدول (۱-۳) مجموعه علائم بکار رفته در این بخش
۶۰	جدول (۲-۳) مقایسه روش های تقاطع طی ۴ بار تکرار
۶۲	جدول (۳-۳) مقایسه روش های جهش
۶۳	جدول (۴-۳) مقایسه نرخ های تعیین شده برای روش تطبیق عملی
۶۴	جدول (۵-۳) انتخاب اندازه جمعیت
۶۹	جدول (۶-۳) فرمول بندی برای دسته های مشتری
۷۶	جدول (۱-۴) مقایسه الگوریتم ها
۸۲	جدول (۲-۴) نمونه داده مولفه های بیرونی
۸۳	جدول (۳-۴) توان مالی و سن مشتریان
۸۶	جدول (۴-۴) محصولات مورد علاقه هر مشتری

«فهرست نمودارها»

۶۱	 نمودار (۱-۳) مقایسه روش های تقاطع
۶۲	 نمودار (۲-۳) مقایسه روش های جهش
۶۳	 نمودار (۳-۳) مقایسه روش های جهش با مقدار شاخص دون
۶۵	 نمودار (۴-۳) بهترین اندازه جمعیت
۷۶	 نمودار (۱-۴) محاسبه فاصله اقلیدوسی برای الگوریتم ها
۷۷	 نمودار (۲-۴) مقدار اعتبار الگوریتم ها در هر بار اجرا
۷۸	 نمودار (۳-۴) میانه و میانگین اعتبار خوشه بندی
۸۰	 نمودار (۴-۴) آزمایش مهاجرت ها با دو روش
۸۴	 نمودار (۵-۴) حاصل خوشه بندی
۸۵	 نمودار (۶-۴) ارزیابی هر خوشه برای هر محصول

«فهرست اشکال»

۱۳	 شکل (۱-۱) منحنی بهینه محلی و بهینه کلی ژنتیک
۱۵	 شکل (۲-۱) مراحل ایجاد الگوریتم
۱۶	 شکل (۳-۱) روشهای ایجاد جمعیت جدید از جمعیت قبلی
۱۷	 شکل (۴-۱) ساختار یک کروموزوم
۱۸	 شکل (۵-۱) شبه کد الگوریتم ژنتیک
۳۱	 شکل (۱-۲) مجموعه داده پروانه ای
۳۲	 شکل (۲-۲) خوشه بندی فازی داده ها
۳۳	 شکل (۳-۲) نحوه ادغام و تفکیک isodata
۵۰	 شکل (۱-۳) مراحل انجام کار در یک نگاه
۷۲	 شکل (۱-۴) خوشه بندی kmeans با نمایش سه بعدی
۷۴	 شکل (۲-۴) خوشه بندی FCM با نمایش سه بعدی
۷۵	 شکل (۳-۴) خوشه بندی ژنتیک پیشنهادی
۷۹	 شکل (۴-۴) نسل تولید الگوریتم پیشنهادی
۷۹	 شکل (۵-۴) نسل تولید الگوریتم ژنتیک مبتنی بر kmean
۸۷	 شکل (۶-۴) قوانین بدست آمده از محصولات

فصل اول

کلیات تحقیق

۱-۱ مقدمه

پیشرفت تکنولوژی تعاملات کسب و کار در سال های اخیر به یک واقعیت تبدیل شده است. تکنولوژی هایی مانند انبار داده ها، داده کاوی و نرم افزار های مدیریت باعث شده اند، مدیریت ارتباط با مشتری را، در منطقه ای جدیدی بستر سازی کنند، تا در آن بتوانند شرکت ها به رقابت بپردازند. از طریق داده کاوی با استخراج اطلاعات پنهان از پایگاه داده های بزرگ، سازمان ها می توانند مشتریان با ارزش را شناسایی و رفتارهای آینده را پیش بینی کنند.[۴]

در بسیاری از شرکت ها، پایگاه اطلاعاتی انبوهی از مشتریان را دارند، که در آنها داده ها از طریق محصولات مورد علاقه مشتریان جمع آوری می شوند. کانال های فروش و مناطق مختلف جغرافیایی وجود دارد که می توان از آنها، برای رسیدن به دانش مطلوب درباره مشتری و تعاملاتی مناسب با مشتریان برقرار نمود، به طوری که محصولات و خدمات ارائه شده به آنها، از مطلوبیت کافی برخوردار و رضایت مشتری را جلب نماید[۵].

داده کاوی تحلیل مرحله ای از کشف دانش در پایگاه داده می باشد[۶]. داده کاوی با بهره گیری از اشتراک مفاهیم موجود در هوش مصنوعی، یادگیری ماشین، آمار، و سیستم های پایگاه به وجود آمده است. هدف کلی از فرایند داده کاوی، استخراج دانش از داده های موجود و تبدیل آن به یک ساختار قابل فهم برای انسان می باشد[۷]. ابزار داده کاوی در حوزه های بسیار متنوع و متفاوت استفاده می شود: حوزه علم: شیمی، فیزیک، داروسازی، تجزیه و تحلیل تصاویر پزشکی، تعیین نوع رفتار با بیماران و پیشگویی میزان موفقیت های اعمال پزشکی، تعیین میزان موفقیت روش های

درمانی در برخورد با بیماری‌های سخت، بیوشیمی، حسگرهای کنترل در اقمار مصنوعی، بیوعلم شامل توصیف ژن‌ها و تقسیم‌بندی گروه پروتئین‌ها و توسعه داروها. خرده فروشی: تجزیه و تحلیل سبد خرید بازار، تعیین الگوهای خرید مشتریان می‌باشند. دو موضوع خوشه بندی و کشف قوانین وابستگی در حوزه داده کاوی بکار برده می‌شود که در این تحقیق، بر روی آنها تمرکز شده است. هدف خوشه بندی داده های مشابه بر روی محیط توزیع شده می باشد، یعنی داده های مشابه در یک دسته قرار می گیرند.

۲-۱ بیان مساله

در تحقیق حاضر، بررسی اینکه به چه میزان دقت، می‌توانیم مشتریان را تقسیم بندی کنیم و از این تقسیم بندی، مشتریان وفادار را استخراج کنیم و همچنین با پیش بینی رفتار خرید محصولات مورد علاقه مشتریان وفادار بتوانیم الگوهای علاقه مندی آنها را به سایر مشتریان ارائه دهیم. داده کاوی که تحلیل داده ها است بعنوان ارتباط بین بخش هایی از داده شناخته می‌شود که می‌تواند به عنوان منبع باارزشی باشد. جمع آوری داده مشتری در داخل یک پایگاه داده واحد، به تمام اجزای درون شرکت (تیم بازاریابی، نیروهای فروش و غیره) این اجازه را می‌دهد تا بتوانند اطلاعات و وظایفشان را با یکدیگر تقسیم کنند.[۸]. در این راستا روش هایی در داده کاوی وجود دارند از جمله طبقه بندی، خوشه بندی، کشف قوانین وابستگی و... که توصیف و پیش بینی داده ها را انجام می‌دهند. الگوریتم های خوشه بندی بسیاری، به منظور دقت در دسته بندی داده ها پیشنهاد شده اند[۹]. در تحقیق حاضر مساله مورد نظر را با دو دید بررسی خواهد شد. یکی از دید زیر ساختی، که با بررسی انواع تکنیک های خوشه بندی و مقایسه بین آنها و ارائه یک راه کار پیشنهادی بر این مسئله، راهی بسوی بهبود بخشیدن دقت خوشه بندی انجام خواهد شد و دیگری از دید کاربردی که بر روی فاکتور های شناسایی رفتار مشتریان، بررسی هایی انجام گرفته خواهد شد، این دید نیز با پیشنهاد شناسایی ویژگی های درونی و بیرونی مشتریان همراه خواهد بود. در اتمام کار، با استفاده از ترکیب موارد پیشنهادی، مطالعه موردی بر روی فروشگاه لوازم خانگی نار بلا انجام گرفته می‌شود.

۱-۳ اهمیت مساله

با بزرگ‌تر شدن حجم داده‌ها و اطلاعات سازمان‌ها، اهمیت این مسئله روزبه‌روز بیشتر می‌شود. تکنولوژی اطلاعات^۱ به طور چشمگیری فرآیندهای کسب و کار را تغییر داده است. اینترنت قدرت چانه زنی مشتریان را نسبت به گذشته بیشتر کرده است. مدیریت ارتباط با مشتری به وجود آمد تا به عنوان یک هدف نهایی برای نگهداشتن مشتری ایفای نقش کند. مدیریت ارتباط با مشتری^۲، از جمله نکات بسیار مهم و حیاتی برای سازمانها به شمار می‌رود و برای پشتیبانی از فرآیندهای بازاریابی، فروش و خدماتی مؤثر، بکار می‌رود. با گسترش اینترنت، ابزارهای در دسترس برای برقراری ارتباط با مشتریان و مدیریت ارتباط با مشتری الکترونیکی^۳ در سازمانها رایج شده‌اند. مشتریان برخط^۴ امروزی، نیازهای مختلف و توقعات بالاتری نسبت به گذشته دارند. مشتریان می‌خواهند خرید کنند و از طریق کانالهای ارتباطی مختلفی چون تلفن، پست الکترونیکی و وب، سرویس دهی شوند. با این قابلیت‌ها، مدیریت ارتباط با مشتری در درون مدیریت ارتباط با مشتری الکترونیکی به تکامل می‌رسد. مدیریت ارتباط با مشتری الکترونیکی جدیدترین پارادایم در دنیای مدیریت ارتباط با مشتری است. شرکتها دیگر نمی‌توانند به استراتژی‌های گذشته بسنده کنند بلکه بایستی همراه با پیشرفت بازار، خودشان را سازگار کنند [۱۰]. تمام سازمان‌هایی که در کسب و کارهای بنگاه به بنگاه^۵ و بنگاه به مشتری^۶ فعالیت می‌کنند بایستی خودشان را در برابر پدیده جدید مدیریت ارتباط با مشتری الکترونیکی سازگار کنند. بر طبق گفته «رومانو»^۷ مدیریت ارتباط با مشتری الکترونیکی جذب و نگهداری اقتصادی مشتریان و ایجاد ارزش برای آنها است. هر روز کسب و کار و خرید مشتریان در اینترنت در حال افزایش است. مشتریان به دلایلی چند، از اینترنت خرید می‌کنند [۱۱]. با مطرح شدن خوشه بندی در داده کاوی چالش‌هایی در خوشه بندی مشتریان به وجود آمده است. روش‌ها و تکنیک‌هایی خوشه بندی بر مبنای kmean مطرح شده است. حساسیت در

^۱ Information Technology

^۲ Customer Relationship Management (CRM)

^۳ Electronic Customer Relationship Management (e-CRM)

^۴ On- Line

^۵ B to B

^۶ B to C

^۷ Romano

تعیین نقاط ابتدایی سرخوشه و همگرایی محلی معمولاً مشکلات موجود در روش های kmean می باشد [۱۲].

با تعریف تکنیک خوشه بندی پیشنهادی، سعی در پیدا کردن مشتریان هم گروه که از نظر ویژگی های درونی و رفتاری مشابهی دارند، را دسته بندی می کنیم، و از توصیف بدست آمده از داده ها می توانند محصولات مورد نظرشان را به مشتریان ارائه دهند. این کار باعث می شود تا مشتریان نسبت به تعاملات آنلاین از پایداری بیشتری برخوردار باشند.

۴-۱ اهداف کلی و اختصاصی

در این تحقیق هدف جستجوی مشتریان وفادار بر اساس الگوریتم خوشه بندی مبتنی بر ژنتیک بهینه خودکار پیشنهادی می باشد و همچنین با پیشنهاد استراتژی مفاهیم مولفه های درونی و بیرونی مشتریان، رفتار مشتریان دسته بندی می شوند. پس از پیشنهاد الگوریتم خوشه بندی و پیاده سازی آن با نرم افزار متلب، با استفاده از ابزار orange قرار است بر روی داده های محصول کشف قوانین انجام گیرد. تا ارائه دقیقی از محصولات مورد علاقه مشتریان را داشته باشیم. که با این کار دقت در خوشه بندی چه در زیر ساختی مساله و چه در کاربرد مساله را، حل می کنیم.

۵-۱ سوال تحقیق

در تحقیق حاضر سوال به این شکل است که آیا می توان رفتار مشتریان را با دقت کامل، با استفاده از الگوریتم خوشه بندی مبتنی بر ژنتیک بهینه پیشنهادی و مفاهیم مولفه های درونی و بیرونی پیشنهادی، پیش بینی کرد. تا بتوانیم ارائه دقیقی از محصولات به مشتریان را داشته باشیم.

۶-۱ تعاریف تحقیق

۱-۶-۱ مدیریت دانش

امروزه در عصر یادگیری، کارکنان از ارزشمندترین منابع سازمان به‌شمار می‌آورند. کارکنان حجم عظیم داده‌ها، اطلاعات و دانش سازمان را تولید کرده و به کار می‌بندند. از این‌رو در سازمان‌های نوین، همکاری مدیریت دانش و مدیریت منابع انسانی، به امری ضروری تبدیل شده است. مدیریت دانش رویکردی سیستمی و یکپارچه است که تمامی دارایی‌های اطلاعاتی سازمان، اعم از: مستندات، پایگاه داده‌ها، خطی‌مشی‌ها، رویه‌ها و تجربیات کارکنان و سازمان را در برمی‌گیرد و به اشتراک می‌گذارد. مدیریت دانش، سکوی تسهیل‌کننده استخراج، ذخیره‌سازی، یکپارچه‌سازی، انتقال، مشاهده، تجزیه و تحلیل و استفاده از دانش در سازمان است. مدیریت دانش به منظور حفظ، تجزیه و تحلیل، سازماندهی، بهبود و تسهیم تجارب کسب و کار، به‌عنوان ابزار، روش و راهبرد در سازمان به‌کار می‌رود. در واقع، مدیریت دانش متدولوژی استخراج سیستمی و استفاده از دانش در سازمان است که از طریق همین دانش و سرمایه فکری، ثروت و ارزش می‌آفریند. در نگرش سیستمی به مدیریت دانش، سه مولفه اصلی: منابع انسانی، فرایندها و فن‌آوری دخالت و با یکدیگر همپوشانی دارند [۱۳].

۲-۶-۱ داده کاوی

پیشینه طرح موضوع داده‌کاوی به دهه ۱۹۸۰ و به صورت جدی، به دهه ۱۹۹۰ برمی‌گردد. پیش از آن از سیستم‌های جمع‌آوری و مدیریت داده‌ها و اصطلاحاً لایروبی داده‌ها استفاده می‌شد، اما به مرور زمان، استخراج و کشف سریع و دقیق اطلاعات با ارزش و پنهان از پایگاه داده‌ها، به عنوان داده‌کاوی مورد توجه قرار گرفت. به این شکل بود که فرایند داده‌کاوی به عنوان فرایند آماری و تجزیه و تحلیل در فرایند کشف دانش در پایگاه داده‌ها (KDD) پررنگ شد، به حدی که گاه،

داده‌کاوی (DM) به عنوان مترادف کشف دانش در پایگاه داده‌ها (KDD) مورد استفاده قرار می‌گرفت [۱۳]. امروزه فرآیند استخراج اطلاعات معتبر، از پیش ناشناخته، قابل فهم و قابل اعتماد از پایگاه داده‌های بزرگ و استفاده از آن در تصمیم‌گیری و در فعالیت‌های تجاری «داده‌کاوی» نامیده می‌شود [۱۴]. در تعاریف متعدد و متنوع برای داده‌کاوی بر موضوعاتی نظیر: استخراج دانش کلان، کاوش در داده‌ها، تجزیه و تحلیل داده‌ها و یافتن روابط و الگوهای مطمئن بین داده‌ها تاکید می‌شود. هدف نهایی داده‌کاوی، ایجاد سیستم‌های پشتیبانی تصمیم‌گیری سازمانی است. داده‌کاوی، به استخراج اطلاعات مفید و دانش از حجم زیاد داده‌ها می‌پردازد. داده‌کاوی، الگوهای حاوی اطلاعات را در داده‌های موجود جست‌وجو می‌کند. این الگوها و الگوریتم‌ها، می‌توانند توصیفی باشند، یعنی داده‌ها را توصیف کنند و یا جنبه پیش‌بینی داشته باشند، یعنی متغیرها برای پیش‌بینی ارزش‌های ناشناخته سایر متغیرها به‌کار روند. داده‌کاوی توصیفی، به دنبال یافتن اگرها در فعالیت‌ها یا اقدامات گذشته است و داده‌کاوی پیش‌بینانه با نگاه به سابقه، رفتار آینده را پیش‌بینی می‌کند [۱۴].

۳-۶-۱ دورنما

راهنمای استخراج الگو از داده‌ها که در این قرن اتفاق افتاده است. در ابتدا با پیدایش روش‌های شناسایی الگوهای داده‌ها تئوری بایس^۸ و تحلیل رگرسیون^۹ آغاز شد. تکثیر، حضور در همه جا و افزایش توان تکنولوژی کامپیوتر که با مجموعه‌ای از داده و ذخیره‌سازی افزایش یافته است. مجموعه‌ای از داده‌ها در اندازه‌ها و پیچیدگی‌هایی رشد یافته‌اند. فرآیند مستقیم افزایش داده‌ها با فرآیند غیر مستقیم، اتوماتیکی پردازش داده‌ها توسعه داده بود. مثل: شبکه‌های عصبی، خوشه‌بندی، الگوریتم ژنتیک، درخت تصمیم‌گیری و پشتیبانی مکانیزم برداری که داده‌کاوی این فرآیندها را در کشف الگوهای پنهان از داده‌ها را برای داده‌های بزرگ بکار می‌گیرد [۱۵] که اینها سالهاست که در حوزه‌های تجاری و دولتی و غیره... در داده‌های بزرگ واریسی می‌کنند. علت اصلی استفاده از داده‌کاوی آنالیز کردن داده‌ها می‌باشد. داده‌کاوی سعی در تلاش برای استاندارد سازی می‌باشد. برای نمونه در سال ۱۹۹۹ در اروپا فرآیند استاندارد کاوش داده‌ها (CRISP-DM 1.0) و در سال

^۸ bayes theorem

^۹ regression analysis

۲۰۰۴، داده کاوی جاوا که برای استاندارد سازی گسترش یافت. نسخه های بعدی از این استاندارد در دست توسعه است. نرم افزار متن باز این سیستم ها به طور آزاد در دسترس می باشد مانند: R project و KNIME، Weka و Rapid miner که برای تعریف فرمال استاندارد سازی تعریف شده اند. با توجه که این سیستم ها توانایی ورود و خروج مدل ها در PMML¹⁰ دارند. برای مدل های داده کاوی روش های استاندارد را تعریف می کند. به منظور اینکه بین برنامه های آماری مختلف اشتراک شده بود. [۱۶] PMML یک زبان پایه ریزی شده بر مبنای XML¹¹ است که توسط گروه داده کاوی توسعه داده شده است. DMG¹² گروه های مستقلی از شرکت های داده کاوی تشکیل شده اند که نسخه PMML در سال ۲۰۰۹ انتشار شده بود. [۱۷، ۱۸، ۱۹]

۷-۱ مفاهیم خوشه بندی

۷-۱-۱ تعاریف

قابلیت خوشه بندی وظیفه تقسیم یک گروه ناهمجنس را در چندین زیر گروه بر عهده دارد. این فرایند یک تفاوت اساسی با طبقه بندی دارد. زیرا در این مدل هیچ گونه الگوی آموزشی نداریم. خوشه بندی به طور خودکار ویژگی های متمایز کننده زیر گروه ها را تعریف می کند و زیر گروه ها را سازماندهی می نماید [۹]. و به عنوان نوعی قابلیت داده کاوی غیر مستقیم مطرح است. این ابزارها پایگاه داده را بر اساس ویژگی های داده ها به چندین بخش تقسیم می کنند و گروه هایی از رکوردها را به وجود می آورند که نمایانگر یا صاحب صفت خاصی هستند. الگوهای به دست آمده در ذات پایگاه داده نهاده شده هستند و نشانگر بعضی اطلاعات غیر منتظره و در عین حال ارزشمند شرکتی باشند [۹]. مثالی از کاربردهای خوشه بندی در تقسیم بندی افرادی است که به پرسش نامه ای پاسخ داده اند. از این کار می توان در تقسیم بندی مشتریانی که به پرسش نامه ها پاسخ داده اند در گروه هایی که اعضای آن بیشترین شباهت را با یکدیگر و بیشترین تفاوت را با

¹⁰ Predictive Model Markup Language

¹¹ Extensible Markup Language

¹² Data Mining Group

اعضای سایر گروه ها دارند، استفاده کرد. بعد از بخش بندی جمعیتی با استفاده از قابلیت خوشه بندی می توان بر روی خوشه های مشخص شده تحلیل هم پیوندی انجام داد تا خریدهای به هم مرتبط یک گروه جمعیتی خاص شناخته شوند. کاربرد خوشه بندی برای تعیین بهترین گروه های جامعه شناختی برای اهداف خاص بازاریابی است. اغلب از خوشه بندی به عنوان اولین گام فرایندهای داده کاری یاد می شود که قبل از سایر فرایندها برای شناسایی گروهی از رکوردهای مرتبط با هم که بعداً بتوانند نقطه آغاز تحلیلها باشند بر روی رکوردها اعمال می شود. هدف از خوشه بندی این است که داده های موجود را به چند گروه تقسیم کنند و در این تقسیم بندی داده های گروه های مختلف باید حداکثر تفاوت ممکن را به هم داشته باشند و داده های موجود در یک گروه باید بسیار به هم شبیه باشند. برخلاف طبقه بندی در خوشه بندی، گروه ها از قبل مشخص نمی باشند و همچنین معلوم نیست که بر حسب کدام خصوصیات گروه بندی صورت می گیرد. در نتیجه پس از انجام خوشه بندی باید یک فرد خبره خوشه های ایجاد شده را تفسیر کند و در بعضی مواقع لازم است که پس از بررسی خوشه ها بعضی از پارامترهایی که در خوشه بندی در نظر گرفته شده اند ولی بی ربط بوده یا اهمیت چندانی ندارند حذف شده و جریان خوشه بندی از اول صورت گیرد. پس از اینکه داده ها به چند گروه منطقی و توجیه پذیر تقسیم شدند از این تقسیم بندی می توان برای کسب اطلاعات در مورد داده ها یا تقسیم داده ها جدید استفاده کنیم.

۱-۷-۲ شرح مختصر تاریخچه تکنیک خوشه بندی

یک رابطه تنگاتنگ بین تکنیک خوشه بندی و دیگر علوم و زمینه های تحقیقاتی وجود دارد، خوشه بندی همیشه در آمار و علوم استفاده می شود. کاربرد های نوعی در این زمینه عبارتند از: پردازش سیگنال صحبت و تشخیص خودکار حروف. الگوریتم های خوشه بندی موجود در یادگیری ماشین، در قطعه بندی تصاویر و بینایی ماشین استفاده شده. مساله خوشه بندی را به عنوان یک مساله تخمین چگالی می توان در نظر گرفت. این نگرش، موضوع تخمین های آماری با توزیع مرکب می باشد. خوشه بندی همچنین بطور گسترده ای در پردازش تصویر به کار رفته، که به معروف است.

۳-۷-۱ تکنیک خوشه بندی

خوشه بندی یعنی، تقسیم داده ها به گروههایی از اشیا مشابه. نمایش داده ها با تعداد کمتر خوشه ها، به الزام باعث ازدست رفتن جزئیات خواهد شد، اما به نوعی ساده سازی به همراه دارد. این تکنیک داده ها را به وسیله خوشه ها مدل می کند. از دیدگاه مدلسازی داده ها، ریشه های خوشه بندی در ریاضیات، آمار و آنالیز عددی می باشد [۲۰]. از دیدگاه یادگیری ماشین، خوشه ها مترادف با الگو های مخفی، جستجو برای یافتن خوشه ها یعنی یادگیری بی سرپرست و نتایج ایجاد شده بیانگر یک مفهوم داده است [۲۱]. از دیدگاه عملی خوشه بندی نقش مهمی در کاربرد های داده کاوی ایفا می کند، چند، نمونه از این کاربرد ها عبارتند از: کاوش داده های علمی، بازیابی اطلاعات، متن کاوی، کاربرد پایگاه داده های نجومی آنالیز وب، مدیریت رابطه با مشتری، بازاریابی، تشخیص پزشکی، محاسبات زیستی، و بسیاری کاربردهای دیگر. خوشه بندی به عنوان یک موضوع تحقیقاتی در چندین زمینه مطرح می باشد، نظیر: آمار، شناسایی الگو، و یادگیری ماشین. در این گزارش به کاربرد خوشه بندی به عنوان تکنیک داده کاوی، توجه شده و از این دیدگاه، خوشه بندی را مورد بررسی قرار می دهیم. داده کاوی پیچیدگی هایی را به خوشه بندی افزود، این پیچیدگی ها عبارتند از مجموعه های بزرگ داده با ابعاد زیاد و انواع متفاوت. این شرایط نیازهای محاسباتی منحصر به فردی در الگوریتمهای خوشه بندی ایجاد کرده است. الگوریتمهای گوناگونی ارائه شده که این نیازهای محاسباتی را ارضاء کرده و با موفقیت از آنها در مسائل واقعی داده کاوی استفاده شده است. [۹].

۴-۷-۱ علائم نشانه گذاری

از X برای استحکام متن وسادگی اصطلاحات، در این قسمت قراردادهایی را انجام می دهیم. هر مجموعه داده ای نظیر تعدادی نقاط داده ای (یابه عنوان مترادف، اشیا، نمونه ها، موارد، الگوها، سطرها، تراکنش ها) تشکیل شده است. این نمایش می دهیم که فضای صفات است و هر

جزء $i = 1:N$ ، $X_i = (x_{i1}, \dots, x_{idm}) \in A$ نقاط داده ای را به صورت یک صفت عددی یا صفت لفظی (یا به طور مترادف، ویژگی، متغیر، بعد، جزء، فضا) می باشد. برای مطالعه بیشتر در مورد نوع داده ای صفات را ببینید. این قالب نقطه-صفت برای داده ها، به طور مفهومی یک را تداعی می کند و در بسیاری از الگوریتم ها از این خاصیت بهره برداری شده است. به هر حال انواع $d \times N$ فضای ماتریس داده ای دیگر نظیر، داده های ترتیبی با طول متغیر و داده های نامرتب، هر چه بیشتر عمومیت پیدا کرده اند. ساده ترین نمایش زیر مجموعه ای از فضای صفات، به صورت ضرب مستقیم کارتزین آن زیر محدوده می باشد $C = \prod_{t=1:d} C_t \subset A$ ، $C_1 \subset A_1$ ، $t = 1:d$ به این زیر فضا، بخش (یا مکعب، سلول، ناحیه) می گوئیم. یک واحد، یک بخش اصلی است که زیر محدوده آن، شامل مقادیر یک دسته باشد، و یا شامل یک فاصله عددی کوچک باشد. توصیف تعداد نقاط داده ای موجود در هر واحد نمایش دهنده یک خوشه بندی افراطی است، یعنی یک هیستوگرام، که در آن هیچ خوشه بندی واقعی انجام نشده است. این نحوه نمایش بسیار پر هزینه می باشد و به لحاظ بیان اطلاعات، از قدرت زیادی برخوردار نیست. بخش بندی با استفاده از دانش کاربر، روش متداولی است که در عمل برای جستجوی داده ها به کار می رود، این روش از دانش فرد خبره بجای تاکید بر یک زیر-قلمرو، استفاده می کند. مابین خوشه بندی و بخش بندی تمایز قائل می شود تا روی اهمیت مساله فرایند یادگیری خودکار، تاکید کنیم. زیرمجموعه یا خوشه است. به طور معمول این زیر k هدف نهایی خوشه بندی تخصیص نقاط به یک سیستم متناهی از مجموعه ها با هم اشتراکی ندارند (این فرض در بعضی اوقات نقض می شود)، و اجتماع آنها با کل مجموعه داده ای برابر است، البته داده های منفرد نیز باید در نظر گرفته شود [9].

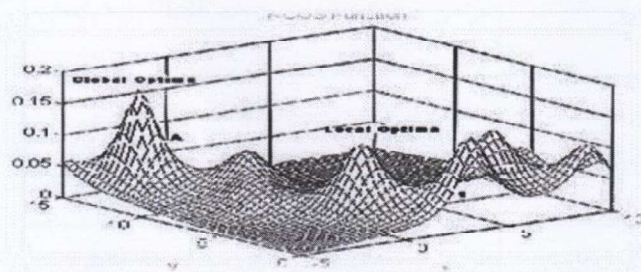
$$X = c_1 Y \dots Y c_k Y_{ck} Y_{c_{outlier}} c_{i1} I c_{i2} \quad (1-1)$$

۸-۱ جستجوی فراکتشافی ژنتیک

۸-۱-۱ مقدمه

الگوریتم های ژنتیک از اصول انتخاب طبیعی داروین برای یافتن فرمول بهینه جهت پیش بینی یا تطبیق الگو استفاده می کنند [۲۲]. الگوریتم های ژنتیک اغلب گزینه خوبی برای تکنیک های پیش بینی بر مبنای رگرسیون هستند. بسیاری از الگوریتم ژنتیک استاندارد برای مدل سازی مسائل مورد استفاده قرار گرفته است [۲۳]. مختصراً گفته می شود که الگوریتم ژنتیک (یا GA) یک تکنیک برنامه نویسی است که از تکامل ژنتیکی به عنوان یک الگوی حل مسئله استفاده می کند. مسئله ای که باید حل شود ورودی است و راه حلها طبق یک الگو کد گذاری می شود و متریک که تابع fitness هم نام دارد هر راه حل کاندید را ارزیابی می کند که اکثر آنها به صورت تصادفی انتخاب می شوند. تابع ارزیابی بر پایه رتبه بندی رشته ای در جمعیت، یا روش هایی از جمله انتخاب تورنمنت [۲۴] مورد استفاده قرار می گیرد. الگوریتم ژنتیک حداقل دارای عناصر: جمعیت کروموزوم، انتخاب بر طبق تابع ارزیابی، تقاطع برای تولید فرزندان، جهش در فرزندان می باشد [۲۵]. هنگامی که لغت تنازع بقا به کار می رود اغلب بار ارزشی منفی آن به ذهن می آید. شاید همزمان قانون جنگل به ذهن برسد و حکم بقای قوی تر! البته برای آنکه خیالتان راحت شود می توانید فکر کنید که همیشه هم قوی ترین ها برنده نبوده اند. مثلاً دایناسورها با وجود جثه عظیم و قوی تر بودن در طی روندی کاملاً طبیعی بازی بقا و ادامه نسل را واگذار کردند در حالی که موجوداتی بسیار ضعیف تر از آنها حیات خویش را ادامه دادند. ظاهراً طبیعت بهترین ها را تنها بر اساس هیكل انتخاب نمی کند! در واقع درست تر آنست که بگوییم طبیعت مناسب ترین ها (Fitness) را انتخاب می کند نه بهترین ها. قانون انتخاب طبیعی بدین صورت است که تنها گونه هایی از یک جمعیت ادامه نسل می دهند که بهترین خصوصیات را داشته باشند و آنهایی که این خصوصیات را نداشته باشند به تدریج و در طی زمان از بین می روند. الگوریتم ژنتیک به دنبال یافتن پاسخ بهینه مسائل مهندسی است. بعنوان مثال اینکه چگونه یک موتور را طراحی کنیم تا بهترین بازدهی را داشته باشد یا چگونه بازوهای یک ربات را متحرک کنیم تا کوتاه ترین مسیر را تا مقصد طی کند. روش های کلاسیک ریاضیات دارای دو اشکال اساسی هستند. اغلب این روش ها نقطه بهینه

محلی^{۱۳} را بعنوان نقطه بهینه کلی^{۱۴} در نظر می‌گیرند و نیز هر یک از این روش‌ها تنها برای مسأله خاصی کاربرد دارند. این دو نکته را با مثال‌های ساده‌ای روشن می‌کنیم. در شکل (۱-۱) منحنی نشان داده شده است، که این منحنی دارای دو نقطه ماکزیمم می‌باشد. یکی از آنها تنها ماکزیمم محلی است. حال اگر از روش‌های بهینه‌سازی ریاضی استفاده کنیم مجبوریم تا در یک بازه بسیار کوچک مقدار ماکزیمم تابع را بیابیم. مثلاً از نقطه ۱ شروع کنیم و تابع را ماکزیمم کنیم. بدیهی است اگر از نقطه ۱ شروع کنیم تنها به مقدار ماکزیمم محلی دست خواهیم یافت و الگوریتم ما پس از آن متوقف خواهد شد. اما در روش‌های هوشمند خاصه الگوریتم ژنتیک بدلیل خصلت تصادفی آنها حتی اگر هم از نقطه ۱ شروع کنیم باز ممکن است در میان راه نقطه A به صورت تصادفی انتخاب شود که در این صورت ما شانس دست‌یابی به نقطه بهینه کلی را خواهیم داشت. در مورد نکته دوم باید بگوییم که روش‌های ریاضی بهینه‌سازی اغلب منجر به یک فرمول یا دستورالعمل خاص برای حل هر مسئله می‌شوند. در حالی که روش‌های هوشمند دستورالعمل‌هایی هستند که به صورت کلی می‌توانند در حل هر مسئله‌ای به کار گرفته شوند [۲۲، ۲۶، ۲۷، ۲۸].



شکل (۱-۱) منحنی بهینه محلی و بهینه کلی ژنتیک

^{۱۳} Local Optima

^{۱۴} Global Optima

۲-۸-۱ الگوریتم ژنتیک چیست؟

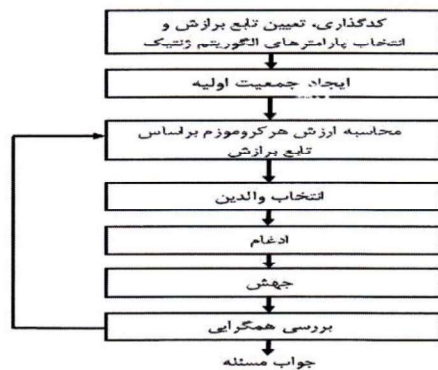
الگوریتمهای ژنتیک یکی از قویترین روشهای برگرفته از طبیعت است که به جستجوی فضای مسأله بصورت تصادفی هدایت شده پرداخته که این جستجو در قالب تلاش جهت ایجاد جوابهایی بهتر، در هر نسل نسبت به جوابهای نسل قبل صورت می گیرد. و یکی از بهترین اشکال بهینه سازی عددی در مسائل علوم و مهندسی را ارائه می کند [۳]. برخی از زمینه های کاربردی الگوریتم ژنتیک بیان شده در [۲] در جدول (۱-۱) نشان داده شده است.

جدول (۱-۱): کاربردهای الگوریتم ژنتیک

زمینه	کاربرد
کنترل	خطوط انتقال گاز - پرتاب موشک - سیستم های تعادلی - رهگیری و ...
طراحی	طراحی هواپیما - طراحی مدارات VLSI و شبکه های ارتباطی و ...
مدیریت و برنامه ریزی	برنامه ریزی تولید - زمانبندی - تخصیص منابع و ...
روبایک	مسیر حرکت روبات و ...
یادگیری ماشین	طراحی شبکه های عصبی - الگوریتم های طبقه بندی و ...
پردازش سیگنال	طراحی فیلتر و ...
سایر موارد	هنر و موسیقی - مسئله فروشنده دورگرد - مسیریابی در شبکه و ...

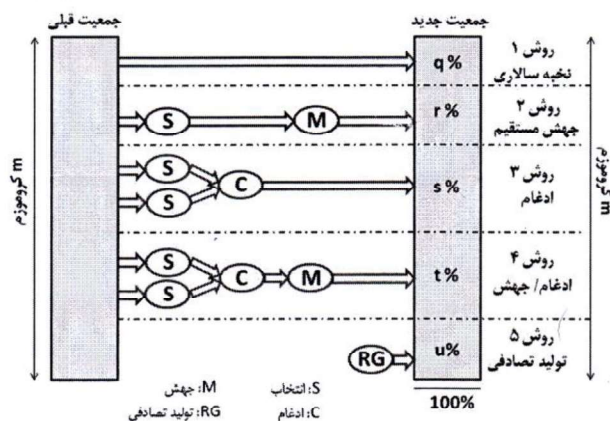
ویژگی اساسی الگوریتم ژنتیک ساده بودن آن می باشد. مراحل الگوریتم ژنتیک بر اساس [۲۹] در شکل ۲-۱ نشان داده شده است. ابتدا پاسخ مسئله را در قالب یک ساختار کروموزمی تعریف می-

نماییم. با معرفی تابع برازش، کیفیت پاسخهای ارائه شده در هر کروموزم را به صورت عدد بیان می- نماییم. سپس تعداد معینی کروموزم به صورت تصادفی (یا شبه تصادفی) تولید می-نماییم که این کروموزمها به عنوان جمعیت اولیه شناخته می شوند. در این مرحله ما تعدادی پاسخ برای مسئله داریم که عمدتاً از کیفیت پایینی برخوردار هستند. میزان کیفیت هر کروموزم از جمعیت را بر اساس تابع برازش مشخص می کنیم. حال با استفاده از روش مناسب (روشی که در آن احتمال انتخاب کروموزم با مقدار برازش بهتر بیشتر از کروموزم دارای مقدار برازش ضعیفتر باشد) دو کروموزم جهت تولید مثل انتخاب می کنیم. سپس با استفاده از این دو کروموزم، کروموزم جدیدی ایجاد می نماییم (جفت گیری). تکامل از یک مجموعه کاملاً تصادفی از موجودیت ها شروع می شود و در نسلهای بعدی تکرار می شود. در هر نسل، مناسبترین ها انتخاب می شوند نه بهترین ها [۲۲، ۲۶، ۲۷، ۲۸] . با یک احتمال مشخص تعدادی از ژنهای بعضی از کروموزمها را تغییر می- دهیم. انجام مراحل انتخاب، جفت گیری و جهش باعث ایجاد جمعیت جدید (نسل جدید) از کروموزمها می گردد. در صورت همگرایی کروموزمها به پاسخ مطلوب عملیات تولید نسل متوقف می گردد. در غیر این صورت ایجاد هر نسل از نسل قبلی تا رسیدن به جواب مطلوب یا برقراری شرط پایان الگوریتم ادامه می یابد.



شکل (۲-۱) مراحل الگوریتم ژنتیک

وجود تنوع زیاد در معرفی فرایند مورد نیاز در هر مرحله باعث ایجاد الگوریتمهای ژنتیک متفاوت می گردد که می تواند به حل مسائل مختلف یا مسائل یکسان با پارامترهای مختلف کمک نماید. لازم به ذکر است که مراحل ذکر شده عنوان فازهای اصلی الگوریتم ژنتیک می باشند و معرفی مراحل جدید بین مراحل ذکر شده امکان پذیر است. به عنوان مثال در [۳۰] نحوه ایجاد جمعیت جدید از جمعیت قدیمی به شکل دقیق تری نشان داده شده است شکل (۲-۱). در شکل (۳-۱) پنج روش برای ایجاد کروموزم های نسل جدید وجود دارد که هر کدام درصدی از کروموزم ها را تولید می نمایند. روش ۱ بر اساس یک تکنیک نخیه گرایی، تعدادی از بهترین کروموزمهای جمعیت قدیمی را بدون تغییر به جمعیت جدید منتقل می نماید. در روش ۲، بر اساس تکنیکهای انتخاب، یک کروموزم انتخاب شده و پس از انجام عملیات جهش بر روی آن به جمعیت جدید افزوده می گردد. روش ۳ و ۴ همان روش مطرح شده در مراحل الگوریتم ژنتیک می باشد. به دلیل احتمالی بودن رخداد جهش، اگر جهش رخ ندهد روش ۳ و اگر رخ دهد روش ۴ مورد استفاده قرار می گیرد. روش ۵ به تولید کروموزم به صورت تصادفی می پردازد. معمولاً برای تولید کروموزمهای نسل جدید از روش ۲ و ۵ استفاده نمی شود (درصد استفاده از آنها صفر می باشد).

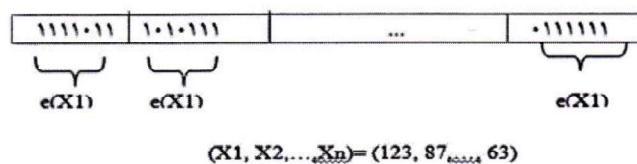


شکل (۳-۱) روشهای ایجاد جمعیت جدید از جمعیت قبلی

از آنجایی که کارایی الگوریتم ژنتیک به شدت وابسته به پارامترهای آن می‌باشد معرفی یک الگوریتم ژنتیک برای گونه‌های مختلف یک مسئله به شدت وابسته به ثابت بودن پارامترهای اساسی آن مسئله می‌باشد [۳۱].

۳-۸-۱ نمونه کروموزم

هر ورودی x در یک عدد برداری $X=(x_1, x_2, \dots, x_n)$ قرار دارد. برای اجرای الگوریتم ژنتیک باید هر ورودی را به یک کروموزوم تبدیل کنیم. می‌توانیم این را با داشتن $\log(n)$ بیت برای هر عنصر و تبدیل ارزش x_i انجام دهیم. شکل (۴-۱) را مشاهده کنید.



شکل (۴-۱) ساختار یک کروموزم

می‌توانیم از هر روش کد کردن برای اعداد استفاده کنیم. در دوره ۰، یک دسته از ورودی‌های X را به صورت تصادفی انتخاب می‌کنیم. بعد برای هر دوره Δm ، ارزش مقدار $Fitness$ تولید، جهش و انتخاب را اعمال می‌کنیم. الگوریتم وقتی پایان می‌یابد که به معیارمان برسیم. سوال مهمی که در اینجا وجود دارد، این است که چه زمانی باید الگوریتم متوقف شود؟

باید اذعان کرد که هیچ جواب جامع و مناسبی برای این سوال وجود ندارد. اما بعضی از شرط‌های توقف عبارتند از:

- ۱- رسیدن به جوابی که این ساده ترین و کوتاه فکرانه ترین روش است. به این معنی که اگر مقدار کروموزم مناسبترین باشد الگوریتم متوقف می شود.
- ۲- عدم پیشرفت: یعنی الگوریتم ژنتیک پس از X بار تکرار با همان کروموزم های قبلی ادامه پیدا کند.
- ۳- به روش آماری: اگر مقدار تابع برازش به یک مقدار مشخصی رسید، الگوریتم متوقف شود.
- ۴- تعداد تکرارها: اگر هیچ یک از موارد فوق جواب نداد، شرط توقف را تعداد تکرارهای مشخصی قرار می دهیم.
- ۵- بهینه سازی موضعی: از یک بهینه ساز موضعی استفاده می کنیم که در صورت عدم پیشرفت متوقف شود. [۳۲]. در شکل (۵-۱) شبه کد الگوریتم نشان داده شده است.

Choose initial population

Repeat

Evaluate the individual fitnesses of a certain proportion of the population

Select pairs of best-ranking individuals to reproduce

Apply crossover operator

Apply mutation operator

Until terminating condition

شکل (۵-۱) شبه کد الگوریتم ژنتیک

۹-۱ مدیریت روابط مشتری

۱-۹-۱ مقدمه

بازاریابی ارتباط با مشتری زمانی که بازار رقابتی بوجود آمده یکی از برجسته ترین نمونه های قابل توجه شده است. امروزه مدل های بازاریابی از مرحله محصول مداری به مرحله مشتری محوری تغییر یافته است. مشتریان ارتباط مختلفی را با تولیدکنندگان تقاضا دارند و این نوع تقاضا را نسبت به مدل قدیمی ترجیح می دهند تکنولوژی پایگاه داده جدید مردم را قادر می سازد که اطلاعات و دانش اینکه چه کسی در حال حاضر مشتری می باشد، چه چیزی و در چه زمانی خریداری می کنند و حتی پیش بینی بر اساس رفتارهای وراثتی را انجام می دهد. هم اکنون بیشتر از هر زمان توانایی درک و مدیریت یک ارتباط بسته با مشتری ارائه می شود. شرکت های موفق از اطلاعات مشتریان به طور آگاهانه برای ساخت روابط با آنها استفاده می کنند. بنابراین مشتری خواهان این ارتباط و گسترش آن می شود. تمرکز CRM^{۱۵} توانایی شرکت ها را برای درک نیازهای کنونی مشتریان و این که چه چیزی در گذشته انجام داده اند و برنامه آینده آنها برای بدست آوردن اهدافشان چیست را افزایش می دهد. هدف ارتقا تجربه مشتریان از چگونگی در تعامل بودن با شرکت هاست. که اگر امیدوارانه بود رضایت بیشتری را بوجود می آورند. که منجر به وفاداری مشتری و فروش بیشتر محصولات می شود. پایگاه داده مرکزی در CRM برای همه در داخل شرکت در دسترس می باشد. با در دسترس بودن این پایگاه داده مرکزی، کلیه اعضای شرکت مشتریان را می شناسند و تجربه کافی را به دست می آورند. بنابراین این مشتریان را از دست نخواهد داد. اگر چه هنر حفظ مشتری و تکنولوژی آن در حال افزایش است، این کار ممکن است برای شرکتهایی که آمادگی آن را ندارند از دست دادن پول باشد [۳۳]

۲-۹-۱ تاریخچه مدیریت روابط مشتری

شاید بتوان تاریخچه ظهور مباحث مرتبط به CRM را در سه دوره زیر خلاصه نمود: الف) دوره انقلاب صنعتی (تولید دستی تا تولید انبوه) ابتکارات فورد در بکارگیری روش تولید انبوه به جای روش تولید دستی، یکی از مهمترین شاخص های این دوره می باشد. هر چند تغییر شیوه تولید

¹⁵ Customer relation managment

باعث شد که محدوده انتخاب مشتریان از نظر مشخصه‌های محصول کاهش یابد (نسبت به تولیدات صنایع دستی) اما محصولات تولید شده به روش جدید از قیمت تمام شده پایین تری برخوردار شدند. به عبارتی دیگر در انتخاب روش تولید انبوه از سوی فرد، افزایش کارایی و صرفه اقتصادی مهمترین اهداف پیش بینی شده بودند.

ب) دوره انقلاب کیفیت (تولید انبوه تا بهبود مستمر) این دوره هم‌زمان با ابتکار شرکت‌های ژاپنی مبنی بر بهبود مستمر فزاینده‌ها آغاز شد. این امر به نوبه خود به تولید کم هزینه تر و با کیفیت تر محصولات منجر شد. با مطرح شدن روشهایی نوین مدیریت کیفیت مانند TQM این دوره به اوج خود رسید. اما با افزایش تعداد شرکت‌های حاضر در عرصه رقابتی و گسترش فرهنگ حفظ و بهبود کیفیت محصول (از طریق ابزارهای مختلف کیفیتی) دیگر این مزیت رقابتی برای شرکتها پیشرو و کارساز نبوده و لزوم یافتن راه‌های جدیدی برای حفظ مزیت رقابتی احساس می شد.

ج) دوره انقلاب مشتری (بهبود مستمر تا سفارشی‌سازی انبوه) در این دوره با توجه به افزایش توقع مشتریان، تولید کنندگان ملزم شدند محصولات خود را با هزینه کم، کیفیت بالا و تنوع زیاد تولید کنند. به معنای دیگر تولید کنندگان مجبور بودند توجه خود را از تولید صرف یافتن راه‌هایی برای رضایت مشتریان سابق خود معطوف نمایند. [۱]

۳-۴-۱ ویژگی‌ها کاربردی CRM

- تمرکز بر مستحکم کردن ارتباط نزدیک تر و عمیق تر با مشتریان
- سودمندی بیشتر مشتریان فعالی نسبت به مشتریان جدید
- تجزیه و تحلیل اطلاعات مشتری برای تصمیم گیری تجاری
- ارتباطات مؤثر با مشتری بر مبنای داده های تبدیل شده به اطلاعات
- بازاریابی اینترنتی، بازاریابی فرد به فرد و بازاریابی از طریق پایگاه داده ها

۴-۴-۱ ضرورت به کارگیری CRM

به این CRM با توجه به اهمیت مشتری به عنوان یکی از ارکان حیات یک سازمان و تأکید

در یک سازمان با CRM عنصر، دلایل زیر را می توان به عنوان ضرورت های استفاده از توجه به پیچیدگی و گستردگی فعالیت های هر سازمان را مطرح کرد:

۱. بهبود خدمات
۲. رضایت مشتری
۳. کاهش هزینه ها
۴. ارتباط فرد به فرد حتی با میلیون ها مشتری.

۵-۹-۱ انواع مشتریان

در بسیاری از موارد از دید سازمان ها مشتریان خوب، کسانی هستند که سهم زیادی در سودآوری سازمان داشته باشند؛ اما در دوران حاضر نمی توان سودآوری و درآمدزایی را تنها معیار تعریف مشتریان اسمی دانست.

شاید بتوان به دو روش مشتریان را دسته بندی کرد:

الف روش جغرافیایی: که با استفاده از روش تجزیه و تحلیل پارتو، مشتریان را مورد بررسی و شناسایی قرار می دهد با در نظر گرفتن این نکته که ۲۰ % مشتریان، تأمین کننده ۸۰ % درآمدهای شرکت می باشند.

ب (روش روانشناختی: که در آن، شناسایی طرز فکر و رفتار مشتری و تأثیرات متقابل مشتریان مد نظر قرار می گیرد. در این نوع نگرش، چهار نوع مشتری شناسایی می شود [۱]:

- مشتریان وفادار
- مشتریان نسبتاً وفادار
- مشتریان بی وفا
- مشتریان وفادار رقیب

فصل دوم

پیشینه تحقیق

۱-۲ روش های خوشه بندی

۱-۱-۲ خوشه بندی انحصاری^{۱۶}، خوشه بندی هم پوشی^{۱۷}، خوشه بندی سلسله مراتبی^{۱۸}،

خوشه بندی مسطح

در روش خوشه بندی انحصاری پس از خوشه بندی هر داده دقیقاً به یک خوشه تعلق می گیرد مانند روش خوشه بندی K-Means. ولی در خوشه بندی با هم پوشی پس از خوشه بندی به هر داده یک درجه تعلق بازاء هر خوشه نسبت داده می شود. به عبارتی یک داده می تواند با نسبت های متفاوتی به چندین خوشه تعلق داشته باشد. نمونه ای از آن خوشه بندی فازی است. در روش خوشه بندی سلسله مراتبی، به خوشه های نهایی بر اساس میزان عمومیت آنها ساختاری سلسله مراتبی نسبت داده می شود. مانند روش تک پیوندی^{۱۹}. ولی در خوشه بندی مسطح تمامی خوشه های نهایی دارای یک میزان عمومیت هستند مانند K-Means. به ساختار سلسله مراتبی حاصل از روش های خوشه بندی سلسله مراتبی دندوگرام^{۲۰} گفته می شود. با توجه با اینکه روش های خوشه بندی سلسله مراتبی اطلاعات بیشتر و دقیق تری تولید می کنند برای تحلیل داده های برای داده های کم پیشنهاد

¹⁶ Exclusive or Hard Clustering

¹⁷ Overlapping or Soft Clustering

¹⁸ Hierarchical

¹⁹ Single Link

²⁰ Dendrogram

می‌شوند ولی از طرفی چون پیچیدگی محاسباتی بالایی دارند برای مجموعه داده‌های بزرگ روش‌های خوشه‌بندی مسطح پیشنهاد می‌شوند. [۹]

۱- روش‌های پایین به بالا^{۲۱} یا متراکم شونده^{۲۲} [۳۴]، [۳۵]: در این روش ابتدا هر داده‌ها به عنوان خوشه‌ای مجزا در نظر گرفته می‌شود و در طی فرایندی تکراری در هر مرحله خوشه‌هایی که شباهت بیشتری با یکدیگر با یکدیگر ترکیب می‌شوند تا در نهایت یک خوشه و یا تعداد مشخصی خوشه حاصل شود. از انواع الگوریتم‌های خوشه‌بندی سلسله مراتبی متراکم شونده رایج می‌توان از الگوریتم‌های تک پیوندی^{۲۳}، میانگین پیوندی^{۲۴} و پیوندی کامل^{۲۵} نام برد [۳۶]، [۳۷]. تفاوت اصلی در بین تمام این روش‌ها به نحوه محاسبه شباهت بین خوشه‌ها مربوط می‌شود. که در بخش‌های بعد به تشریح هر یک پرداخته خواهد شد.

۲- روش‌های بالا به پایین^{۲۶} یا تقسیم کننده^{۲۷} [۳۴]، [۳۵]: در این روش ابتدا تمام داده‌ها به عنوان یک خوشه در نظر گرفته می‌شوند و سپس در طی یک فرایند تکراری در هر مرحله داده‌هایی شباهت کمتری به هم دارند به خوشه‌های مجزایی شکسته می‌شوند و این روال تا رسیدن به خوشه‌هایی که دارای یک عضو هستند ادامه پیدا می‌کند.

۱- خوشه‌بندی با روش تک پیوندی: این روش یکی از قدیمی‌ترین و ساده‌ترین روش‌های خوشه‌بندی است و جزء روش‌های خوشه‌بندی سلسله مراتبی و انحصاری محسوب می‌شود. به این روش خوشه‌بندی، تکنیک نزدیکترین همسایه^{۲۸} نیز گفته می‌شود. در این روش برای محاسبه شباهت بین دو خوشه A و B از معیار (۱-۲) استفاده می‌کند.

21 Bottom-Up
22 Agglomerative
23 Single-Link
24 Average-Link
25 Complete-Link
26 Top-Down
27 Divisive
28 Nearest Neighbour

$$d_{AB} = \min d_{ij} \quad (1-2)$$

که i یک نمونه داده متعلق به خوشه A و j یک نمونه داده متعلق به خوشه B می باشد. در واقع در این روش شباهت بین دو خوشه، کمترین فاصله بین یک عضو از یکی با یک عضو از دیگری است. [۳۸]

۲ - خوشه بندی با روش پیوندی کامل: این روش همانند تک پیوندی جزء روشهای خوشه بندی سلسله مراتبی و انحصاری محسوب می شود. به این روش خوشه بندی، تکنیک دورترین همسایه^{۲۹} نیز گفته می شود. در این روش برای محاسبه شباهت بین دو خوشه A و B از معیار (۲-۲) استفاده می شود:

$$d_{AB} = \max_{i \in A, j \in B} d_{ij} \quad (2-2)$$

که i یک نمونه داده متعلق به خوشه A و j یک نمونه داده متعلق به خوشه B می باشد. در واقع در این روش شباهت بین دو خوشه بیشترین فاصله بین یک عضو از یکی با یک عضو از دیگری است. [۳۸].

۳ - خوشه بندی با روش میانگین پیوندی: این روش همانند تک پیوندی جزء روشهای خوشه بندی سلسله مراتبی و انحصاری محسوب می شود. از آنجا که هر دو روش خوشه بندی تک پیوندی و پیوندی کامل بشدت به نویز حساس می باشد، این روش که محاسبات بیشتری دارد، پیشنهاد شد. در این روش برای محاسبه شباهت بین دو خوشه A و B از معیار (۳-۲) استفاده می شود:

²⁹ furthest Neighbour

$$d_{AB} = \frac{\sum_{i \in A, j \in B} d_{ij}}{N_A N_B} \quad (3-2)$$

که i یک نمونه داده متعلق به خوشه A و j یک نمونه داده متعلق به خوشه B می‌باشد. و N_A تعداد اعضا خوشه A و N_B تعداد اعضا خوشه B است. در واقع در این روش، شباهت بین دو خوشه میانگین فاصله بین تمام اعضا یکی با تمام اعضا دیگری است. [۳۸]

۴ خوشه‌بندی با روش گروه پیوند میانگین^{۳۰}: این روش همانند تک پیوندی جزء روشهای خوشه‌بندی سلسله مراتبی و انحصاری محسوب می‌شود. به این روش فاصله مرکزی^{۳۱} نیز گفته می‌شود. در این روش برای محاسبه شباهت بین دو خوشه A و B از معیار (۴-۲) استفاده می‌شود:

$$d_{AB} = d\left(\frac{\sum_{i \in A} X_i}{N_A}, \frac{\sum_{j \in B} X_j}{N_B}\right) \quad (4-2)$$

که X_i یک نمونه داده متعلق به خوشه A ، X_j یک نمونه داده متعلق به خوشه B ، N_A تعداد اعضا خوشه A و N_B تعداد اعضا خوشه B است. در واقع در این روش، شباهت بین دو خوشه فاصله بین بردار میانگین تمام اعضا یکی با بردار میانگین تمام اعضا دیگری است.

³⁰ Group Average Link

³¹ Centroid Distance

۲-۱-۲ روش خوشه‌بندی K-Means (C-Means یا C-Centeriod)

این روش علی‌رغم سادگی آن یک روش پایه برای بسیاری از روش‌های خوشه‌بندی دیگر (مانند خوشه‌بندی فازی) محسوب می‌شود. این روش، روشی انحصاری و مسطح محسوب می‌شود. برای این الگوریتم شکل‌های مختلفی بیان شده است. ولی همه آنها دارای روالی تکراری هستند که برای تعدادی ثابت از خوشه‌ها سعی در تخمین موارد زیر دارند:

- بدست آوردن نقاطی به عنوان مراکز خوشه‌ها این نقاط در واقع همان میانگین نقاط متعلق به هر خوشه هستند.
- نسبت دادن هر نمونه داده به یک خوشه که آن داده کمترین فاصله تا مرکز آن خوشه را دارا باشد.

در نوع ساده‌ای از این روش ابتدا به تعداد خوشه‌های مورد نیاز نقاطی به صورت تصادفی انتخاب می‌شود. سپس در داده‌ها با توجه با میزان نزدیکی (شباهت) به یکی از این خوشه‌ها نسبت داده می‌شوند و بدین ترتیب خوشه‌های جدیدی حاصل می‌شود. با تکرار همین روال می‌توان در هر تکرار با میانگین‌گیری از داده‌ها مراکز جدیدی برای آنها محاسبه کرد و مجدداً داده‌ها را به خوشه‌های جدید نسبت داد. این روند تا زمانی ادامه پیدا می‌کند که دیگر تغییری در داده‌ها حاصل نشود. [۳۹] فرمول (۲-۵) به عنوان تابع هدف مطرح است

$$J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2 \quad (۵-۲)$$

که $\| \cdot \|$ معیار فاصله بین نقاط و c_j مرکز خوشه j ام است.

الگوریتم زیر الگوریتم پایه برای این روش محسوب می‌شود:

۱. در ابتدا K نقطه به عنوان به نقاط مراکز خوشه‌ها انتخاب می‌شوند.

۲. هر نمونه داده به خوشه‌ای که مرکز آن خوشه کمترین فاصله تا آن داده را داراست، نسبت داده می‌شود.
۳. پس تعلق تمام داده‌ها به یکی از خوشه‌ها برای هر خوشه یک نقطه جدید به عنوان مرکز محاسبه می‌شود. (میانگین نقاط متعلق به هر خوشه)
۴. مراحل ۲ و ۳ تکرار می‌شوند تا زمانی که دیگر هیچ تغییری در مراکز خوشه‌ها حاصل نشود.

۳-۱-۲ مشکلات روش خوشه‌بندی K-Means

علی‌رغم اینکه خاتمه‌پذیری الگوریتم بالا تضمین شده است ولی جواب نهایی آن واحد نبوده و همواره جوابی بهینه نمی‌باشد. به طور کلی روش ساده بالا دارای مشکلات زیر است.

- جواب نهایی به انتخاب خوشه‌های اولیه وابستگی دارد.
 - روالی مشخص برای محاسبه اولیه مراکز خوشه‌ها وجود ندارد.
 - اگر در تکراری از الگوریتم تعداد داده‌های متعلق به خوشه‌ای صفر شد راهی برای تغییر و بهبود ادامه روش وجود ندارد.
 - در این روش فرض شده است که تعداد خوشه‌ها از ابتدا مشخص است. اما معمولاً در کاربردهای زیادی تعداد خوشه‌ها مشخص نمی‌باشد.
- در این تحقیق سعی خواهد شد تا مشکلات انتخاب مرکز خوشه‌ها برای داده‌ها انجام گیرد. که این کار با جاگیری بهترین موقعیت برای دسته بندی انجام گرفته می شود. و همچنین بر مبنای میزان پراکندگی تعداد خوشه‌ها با فرمول بندی مناسب انجام می گیرد.

۳-۱-۲ خوشه بندی با روش وارد^{۳۲}

این روش نیز همانند تک پیوندی جز روش های خوشه بندی سلسله مراتبی و انحصاری محسوب می شود. در این روش خوشه بندی برای کاهش تلفات ناشی داده های دور افتاده از

³² Ward

معیاری جدید برای محاسبه عدم شباهت بین خوشه ها استفاده می کند. در روش واردز از مجموع مربعات تفاضل هر داده از یک خوشه با بردار میانگین آن خوشه به عنوان معیاری برای سنجش یک خوشه استفاده می شود. الگوریتم زیر را می توان برای روش واردز در نظر گرفت. [۴۰]

- ۱- ابتدا هر داده به عنوان یک خوشه در نظر گرفته می شود.
- ۲- به ازای تمام جفت خوشه های ممکن از مجموعه خوشه ها آن دو خوشه ای که مجموع مربعات تفاضل داده های خوشه حاصل از اجتماع آنها با بردار میانگین خوشه حاصل کمینه باشد، انتخاب می شوند.
- ۳- دو خوشه انتخاب شده با هم ترکیب می شوند.
- ۴- تا زمانی که تعداد خوشه ها به تعداد مورد نظر نرسیده است، دو، سه و چهار را تکرار کن.

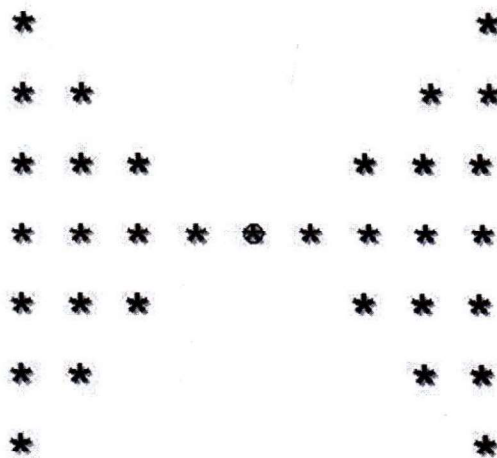
۵-۱-۲ خوشه بندی فازی

روسپینی^{۳۳} اولین مدل خوشه بندی با ایده فازی را مطرح کرد. برای درک بهتر خوشه بندی فازی و الگوریتمهای مختلف آن لازم است تا ابتدا با مفهوم مجموعه های فازی و تفاوت آنها با مجموعه های کلاسیک آشنا شویم. در مجموعه های کلاسیک یک عضو از مجموعه مرجع یا عضوی از مجموعه A است یا عضو مجموعه A نیست. مثلاً مجموعه مرجع اعداد حقیقی را در نظر بگیرید. عدد ۲٫۵ عضو مجموعه اعداد صحیح نمی باشد حال آنکه عدد ۲ عضو این مجموعه است. به زبان دیگر تعلق عدد ۲٫۵ به مجموعه اعداد صحیح ۰ است و تعلق عدد ۲ به این مجموعه ۱ است. در واقع می توان برای هر مجموعه یک تابع تعلق تعریف کرد که مقدار این تابع تعلق برای اعضای مجموعه ۱ می باشد و برای بقیه ۰. در مجموعه های کلاسیک مقدار این تابع تعلق یا ۰ است یا ۱. حال مجموعه انسان های جوان و پیر را در نظر بگیرید. سوالی که در اینجا مطرح می شود این است که آیا فردی با سن ۲۵ جزء این مجموعه است یا خیر؟ سن ۳۰ چطور؟ ۳۵ همانطور که حدس زدید نمی توان بطور قطع و یقین مرزی برای انسان های جوان و پیر در نظر گرفت. دلیل آن هم این است که اگر فرضاً ۳۵ جوان محسوب شود ۳۶ نیز می تواند جوان باشد و همینطور ۳۷ و

³³ Ruspini

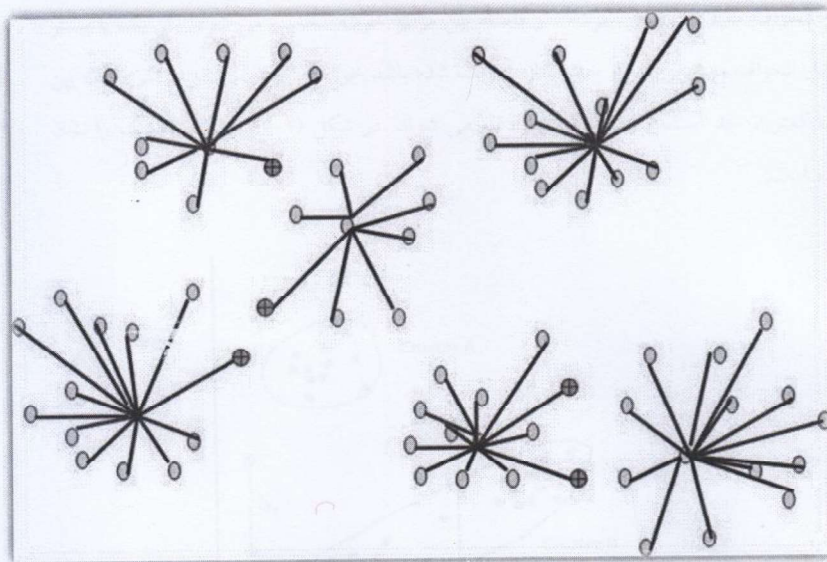
۳۸ و غیره. در واقع در اینجا با مفهوم عدم قطعیت^{۳۴} مواجه هستیم. ما خودمان نیز از عدم قطعیت در زندگی روزمره بارها استفاده کرده ایم مثلاً هوای سرد، آب داغ و غیره. در واقع تمامی مثالهای بالا مثالهایی از مجموعه های فازی می باشند. تفاوت اصلی مجموعه های فازی و مجموعه های کلاسیک در این است که تابع تعلق مجموعه 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 بگیرید اگر ۲۵ سال را سن جوانی در نظر بگیریم می توانیم به ۲۵ تعلق ۱ بدهیم و مثلاً به ۳۰ تعلق ۰٫۸ و به ۳۵ تعلق ۰٫۷۵ و به ۹۰ تعلق ۰٫۱ را بدهیم. اگر اعضای یک مجموعه فازی تنها دارای تابع تعلق ۰ و ۱ باشند این مجموعه فازی یک مجموعه کلاسیک خواهد بود. نکته جالب توجه این است که مثلاً سن ۵۰ می تواند با تعلق ۰٫۵ عضو مجموعه جوان باشد و با تعلق ۰٫۵ عضو مجموعه پیر یعنی یک عضو مجموعه مرجع می تواند با درجه های تعلق مختلف عضو مجموعه های فازی تعریف شده روی مجموعه مرجع باشد. 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. مثلاً در شکل دو 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. برای روشن شدن مطلب شکل (۲-۱) را در نظر بگیرید:

³⁴ Uncertainty



شکل (۱-۲) مجموعه داده پروانه ای

اگر نمونه های ورودی مطابق شکل فوق باشند مشخص است که می توان داده ها را به دو خوشه تقسیم کرد اما مشکلی که پیش می آید این است که داده مشخص شده در وسط می تواند عضو هر دو خوشه باشد بنابراین باید تصمیم گرفت که داده مورد نظر متعلق به کدام خوشه است، خوشه سمت راست یا خوشه سمت چپ. اما اگر از خوشه بندی فازی استفاده کنیم داده مورد نظر با تعلق ۰،۵ عضو خوشه سمت راست و با تعلق مشابه عضو خوشه سمت چپ است. تفاوت دیگر در این است که مثلاً نمونه های ورودی در سمت راست شکل (۱-۲) می توانند با یک درجه تعلق خیلی کم عضو خوشه سمت چپ نیز باشند که همین موضوع برای نمونه های سمت چپ نیز صادق است. بعنوان یک مثال دیگر شکل (۲-۲) را در نظر بگیرید. در این شکل نمونه هایی که با علامت بعلاوه مشخص شده اند به بیش از یک خوشه تعلق دارند [۴۱].



شکل (۲-۲) خوشه بندی فازی داده

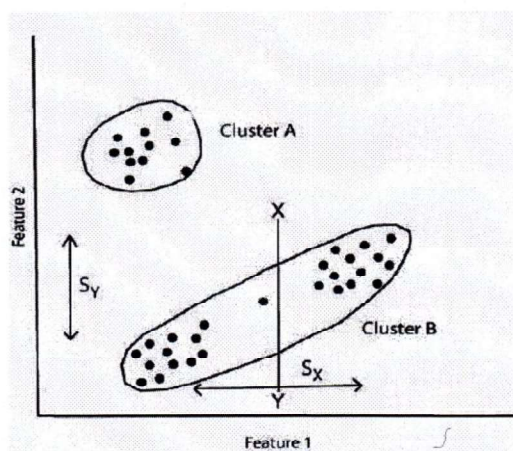
۶-۱-۲ الگوریتم Iso data

روش Iso data روش توسعه یافته توسط توپ، هال و دیگران در سال ۱۹۶۰ می باشد. Iso data روشی در طبقه بندی بدون ناظر می باشد و نیازی به دانستن تعداد خوشه ندارد. این الگوریتم خوشه ها را تفکیک و ادغام می کند. حد آستانه ای برای مقدار پارامتر در نظر گرفته می شود و الگوریتم تا زمان رسیدن به این حد آستانه به کار خود ادامه می دهد [۴۲].

۶-۱-۲ شیوه کار isodata

۱- مراکز به طور تصادفی قرار داده می شوند و نقاط بر مبنای کوتاهترین فاصله به مرکز تخصیص داده می شوند.

۲- انحراف معیار درون هر خوشه ، و فاصله بین مراکز خوشه محاسبه می شود. اگر یک یا بیشتر مقدار انحراف معیار بزرگتر از حد آستانه تعریف شده باشد خوشه ها جدا می شوند. اگر فاصله بین آنها کمتر از حد آستانه باشد خوشه ها ادغام می شوند. در شکل (۳-۲) ادغام و تفکیک را نشان داده است.



شکل (۳-۲) نحوه ادغام و تفکیک isodata

۳- تکرار دوم با مرکز خوشه جدید انجام داده می شود.

۴- تکرار های بعدی انجام می دهند تا زمانیکه: ۱- میانگین فاصله مراکز داخلی زیر حد آستانه تعریف شده باشند. ۲- میانگین تغییرات در فاصله مراکز داخلی بین تکرار ها کمتر از یک حد آستانه باشند. ۳- یا به بیشترین مقدار تکرار دستیابی شود. [۴۲]

۲-۶-۱-۲ مزایا و معایب

۱- معایب Isodata :

ممکن است این روش زمانبر باشد موقعی که داده ها غیر ساختیافته باشد. الگوریتم می تواند به شکل مارپیچی خارج از کنترل، یک طبقه را از دست دهد. [۴۲]

۲- مزایای Isodata :

نیازی به دانستن اطلاعات در مورد داده ها قبل از اجرا را ندارد. تلاش کاربری کمی مورد نیاز است. Isodata در شناسایی خوشه های طیفی بسیار موثر است. [۴۲]

۲-۱-۷ الگوریتم Isoclus

Isoclus یک الگوریتم خوشه بندی بر اساس Isodata است. الگوریتم خوشه بندی با تغییرات جزئی انجام گرفته شده است. Isoclus مانند الگوریتم k-mean سعی می کند تا بهترین مراکز خوشه را انتخاب کند. Isoclus با استفاده از تخمین های متفاوت تعیین می کند که چه موقع خوشه ها از هم جدا یا ادغام شوند. تعدادی پارامتر در زیر آمده است.

۱- تعداد خوشه های مورد نظر (NUM CLUS)

۲- حداقل تعداد نمونه ها در یک خوشه (Samp Rm)

۳- حداقل تعداد تکرار (MaxIter)

۴- حداکثر انحراف معیار در هر خوشه (Std Dev)

۵- پارامتر لامپینک (Lump) و حداکثر تعداد زوج ها که می تواند یکجا در هر تکرار باشند (MaxPair)

این الگوریتم در مجموعه داده های بزرگ به کندی اجرا می شود. در این الگوریتم اولین نقاط نمونه ها به طور تصادفی از یک مجموعه اولیه می باشند. سپس به طور تصادفی تعداد خوشه از این نمونه ها انتخاب می شود. بعد از ورودی مقادیر پارامتر ها در گام اول، گام های بعدی تا زمانی که پایان یابد ادامه می دهد. در گام دوم، فاصله نقاط نسبت به مرکز را محاسبه می کند و نقاط به نزدیکترین مراکزشان تخصیص داده می شوند. و در گام سوم خوشه هایی که از $Samp\ Rm$ کمتر باشند حذف می شوند. و در گام چهارم مراکز خوشه به میانه نمونه ها در خوشه های باقی مانده انتقال داده می شوند. گام دوم تا چهارم ادامه داده می شود تا خوشه ای در گام سوم حذف نشود. در گام دوم و چهارم مشاهده می کنیم که این گام تشکیل یک تکرار از الگوریتم لوید را انجام می دهند. این الگوریتم میانگین فاصله نمونه ها از مراکز خوشه نزدیک را محاسبه میکنند که در گام پنجم قرار دارد و در گام ششم، در نهایت کل میانگین فاصله محاسبه می شود. در گام های بعدی الگوریتم تفکیک و ادغام خوشه ها را در نظر قرار می دهد. در گام چهارم رابطه بین تعداد خوشه های بدست آمده تا کنون و $NUM\ CLUS$ و تعداد تکرار و $MaxIter$. این الگوریتم یک یا چندین گام های زیر را رد کند. گام هشتم برای هر خوشه انحراف معیار در هر یک از محور مختصات بدست می آید. در گام نهم حداکثر مقدار انحراف معیار در امتداد هر یک از مختصات بدست می آورد. در گام دهم بر اساس این اطلاعات و پارامتر های $StdDev$ و $NumClus$ این الگوریتم خوشه های بزرگ با انحراف معیار بالا را تفکیک می کند. در گام یازده تا سیزده اگر الگوریتم جدا باشد الگوریتم به گام دوم بر می گردد و گرنه پارامتر $Lump$ و $MaxPair$ برای زوج خوشه ها مورد استفاده قرار می دهد. در گام چهاردهم تعیین می کند که الگوریتم پایان یافته و یا تعداد تکرار ها افزایش پیدا کند. تفاوت الگوریتم $Isoclus$ با $Isodata$ در گام دوم تا چهارم الگوریتم لوید است. در $Isoclus$ این مراحل تا زمانی که خوشه ای در گام سوم حذف نشود تکرار می شود ولی در حالیکه الگوریتم $Isodata$ این گام ها را فقط یکبار در طول هر تکرار انجام می دهد. [۴۳]

۲-۲ پیشینه جنبه ساختاری مورد مطالعه

اخیراً روش هایی چون مثل الگوریتم GKM پیشنهاد شده است که ایده اصلی استراتژی نگهداری جمعیت می باشد. مجموعه داده مورد بحث خوشه بندی شامل N نقاط است و N الگوهای متناظرشان است. هر الگو یک بردار D بعدی است. هدف استراتژی تقسیم بندی N الگو به K خوشه جاییکه k پارامتر تعریف شده توسط کاربر می باشد، این کار همزمان با حداقل کردن میانگین مجذور خطا انجام می شود. به این معنی که اگر داشته باشیم $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n$ و در آن N الگوی پارامتر باشد و x_{nd} اشاره به d امین بعد ویژگی از الگوی $(n = 1, \dots, N)$ و همچنین G_k اشاره به k امین خوشه، Z_k اشاره به تعداد الگو در G_k می باشد.

سپس $\vec{c}_k (c_{k1}, c_{k2}, c_{k3}, \dots, c_{kn})$ از خوشه G_k می تواند به این صورت

$$C_{kd} = \frac{SF_{kd}}{Z_k} = \frac{\sum_{\vec{x} \in G_k} x_{nd}}{Z_k}, (d = 1, \dots, D)$$

جاییکه SF_{kd} مجموع d امین ویژگی از تمام الگو در G_k می باشد. سپس میانگین مجذور خطا به صورت معیار زیر تعریف می شود.

فرمول مورد استفاده شده در این الگوریتم تابع هدف است این الگوریتم دارای پیچیدگی زمانی برای محاسبه میانگین مجذور خطا $O(N * D)$ می باشد [۴۴].

در الگوریتم FGC ضمن پیدا نمودن تعداد بهینه مرکز خوشه ها با در نظر گرفتن برخی اشیا داده به عنوان مراکز خوشه ها باعث کارایی بهتر الگوریتم خوشه بندی می شود. در این الگوریتم ابتدا ماتریس فاصله ها که اقلیدوسی هر شی داده نسبت به دیگر اشیا داده را نگهداری می کند، محاسبه می شود. سپس بازای هر شی داده i یک شعاع همسایگی r تعریف می شود تا در صورتی که آن شی داده به عنوان مرکز خوشه قرار گیرد، میزان تعلق شی داده i ($i \neq j$) که در فاصله r از شی مرکزی

ز قرار می گیرد، $u_{ij} = 0.5$ شود. مقدار ۲ میزان پراکندگی مجموعه داده ها بستگی دارد. پس در گام بعد، ماتریس درجه عضویت (تعلق) U مطابق رابطه (۶-۲) محاسبه یا بروزرسانی می شود. [۴۵]

$$u_{ij} = \frac{1}{1 + \left(\frac{d_{ij}^2}{\eta_i} \right)^{1/(m-1)}} \quad (6-2)$$

لازمو و موخرجییک الگوریتم ژنتیک برای استخراج مرکز خوشه در الگوریتم k-means ارائه دادند. مجموعه ای از مراکز خوشه با استفاده از فوق-چهار درختی^{۳۵} روی داده ها ساخته شده است. الگوریتم k-means که به طور گسترده ای برای خوشه به خاطر کارایی محاسباتی آن استفاده می شوند، با توجه به N نقطه در فضای D بعدی و تعداد خوشه مورد نظر K ، K (در اینجا به معنی در پی مجموعه ای از خوشه K) مراکز برای به حداقل رساندن مجموع مجذور فاصله اقلیدسی بین هر نقطه و نزدیکترین مرکز خوشه آن است. با این حال، الگوریتم بسیار حساس به انتخاب اولیه از مراکز است و به احتمال زیاد با همگرایی به پارتیشن است که به میزان قابل توجهی پایین تر از بهینه سراسری است. در همین راستا (GA) پیشنهادیشان مربوط به مراکز الگوریتم k-means که به طور همزمان شناسایی پارتیشن های خوبی برای طیف وسیعی از مقادیر در اطراف K مشخص شده است. مجموعه ای از مراکزی که نشان داده با استفاده از فوق چهار درختی ساخته شده بر روی داده ها می باشد. این نمایش در GA پیشنهادیشان باعث استخراج یک جمعیت اولیه از مراکز خوب و حمایت از عملیات تقاطع جدید که به صورت انتخابی زیر مجموعه های خوبی از مراکز همسایه از والدین به فرزندان از طریق مبادله subtrees عبور داده می شوند [۴۶].

³⁵ Hyper-quadtree

الگوریتم ژنتیک AGCUK. این الگوریتم ژنتیک با استفاده از روش نویزینگ مسائل را حل کرده است. و تکنیک های بهینه به وسیله اصول ارزیابی و ژنتیک های طبیعی تعلیم داده شده اند. آنها دارای مقدار بزرگی از مشابهت های ضمنی و راه حل های بهینه نزدیکی از اهداف یا تابع ارزیابی در پیچیدگی، بزرگی و نگاه چند مدلی را دارا می باشند. که این تکنیک فوق تخمینی توسط کوران وهادری پیشنهاد شده است. بجای گرفتن داده های خام برای محاسبه به طور مستقیم، روش نویزینگ با در نظر گرفتن نتایج بهینه به صورت خروجی داده های بی ثبات به طرف داده واقعی حرکت می کند. در الگوریتم AGCUK کروموزم از اعداد واقعی مختصات مرکز خوشه ساخته می شود. طول کروموزم $m \times k_i$ ، جاییکه k_i اشاره به تعداد کلاسترهای i امین اشخاص می کند و m اشاره به تعداد خصوصیات اشیا می کند. در این الگوریتم برای اشخاص i و تعداد k_i خوشه به طور تصادفی در محدوده $[k_{min}, k_{max}]$ تولید می شود در اینجا k_{min} با مقدار ۲ انتخاب می شود، مگر تخصیص داده شده باشد. و برای k_{max} مقدار $\sqrt{N}$ تخصیص داده می شود. در این مقاله، ارزیابی یک شخص با استفاده از شاخص داویز بولدینگ استفاده شده است. و همچنین این روش با دیگر ارزیابی خوشه وفق داده شده است. شاخص DB تابعی از نرخ مجموع پراکنندگی درون خوشه بین خوشه های جدا می باشد [۴۸]. در مقاله دیگر مدل تطبیق مهاجر GA (SAMGA) جایی که پارامتر اندازه جمعیت تعداد نقاط تقاطع و نرخ جهش برای هر جمعیت تنظیم شده است، علاوه بر آن افراد مهاجر بین جمعیت به صورت پویا تصمیم گرفته می شوند. در اینجا یک تحلیل ریاضی با روش کیفی و اینکه الگوریتم نشان می دهد که دانش کشف شده در جایی که جستجو می شود با تخمین بالایی انجام می دهد و این کار را همزمان جستجو در فضای دیگری انجام می دهد و در نهایت کارایی این الگوریتم از توابع استاندارد تست شده است. طرح میشیگان طبقه بندی شده، مورد استفاده قرار گرفته و سیستم با پایگاه داده ماشین یادگیر و غیره تست شده و کارایی الگوریتم شان بهتر از دیگران می باشد. در این الگوریتم چهار مورد پیشنهاد شده است. اولین پارامتر اندازه جمعیت را تغییر می دهد. اندازه جمعیت به طور پویا، تابع ارزیابی بهترین فردی از جمعیتی که با میانه تابع ارزیابی جمعیت مقایسه شده است. با بکارگیری فرمول فوق اندازه جمعیت با تابع ارزیابی

بزرگتر از میانه جمعیت می باشد که این حالت را رشد داده می شود و کوچکتر از آن حذف می شوند. در نتیجه این کار باعث می شود بهترین ناحیه برای جستجو انتخاب شود. پارامتر دوم به طور پویا تعداد نقاط تقاطع را محاسبه می کند با به کار بردن فرمول پیشنهادی توانسته است تعداد افراد در جمعیت کمتر از میانه تعداد افراد در جمعیت را انجام دهد، سپس تعداد نقاط تقاطع افزایش داده می شود. در روش غیرمستقیم بروز رسانی تعداد نقاط تقاطع با تابع ارزیابی جمعیت پیوند داده می شود. از بروز رسانی اندازه جمعیت واضح است که اندازه جمعیت به همراه تابع ارزیابی جمعیت افزایش پیدا می کند اما از بروز رسانی تعداد نقاط تقاطع دیده اند که که بطور نسبی اندازه جمعیت کم ، تعداد نقاط تقاطع افزایش داده می شود. بنابراین جستجو از اکتشاف بیشتری برخوردار خواهد بود. پارامتر سوم نرخ جهش را مشابه با محاسبه تعداد نقاط تقاطع انجام می دهد. که با نرخ های اختیاری تعیین شده است. پارامتر چهارم نرخ مهاجرت را مشخص کرده است. یعنی مهاجرت افراد از یک جمعیت به جمعیت دیگری انجام گرفته است. مهاجرت در کشف طرح جدیدی که بدست آمده از طرح تقاطع دو جمعیت است. این پارامتر به طور دقیق نمی باشد و مهاجرت موقعی اتفاق می افتد که میانگین تابع ارزیابی جمعیت باقی مانده تغییر نشده باشد. وقتی که اندازه جمعیت در وضعیت ثابتی قرار گیرد. رویداد مهاجرت سعی می کند تا طرح جدیدی برای کشف ارائه دهد. این الگوریتم پیشنهادی اکتشاف عمقی از ناحیه امیدبخش بطور تخمینی روی فضای جستجو انجام گرفته است و اکتشاف بهنایی را از ناحیه دیگری روی فضای جستجو انجام داده است [۴۸].

در کار هانگ هی و یانگ هونگ تان ، الگوریتم دو مرحله ژنتیک برای خوشه بندی انجام داده اند. الگوریتم خوشه بندی ژنتیک دو مرحله ای پیشنهاد شده است. این الگوریتم بطور اتوماتیک تعداد مناسبی از خوشه ها و تقسیم مناسب از مجموعه داده ها را مشخص میکند. در مرحله انتخاب و جهش با تغییر دادن احتمالات انتخابی و جهش بر طبق تعداد خوشه ها در جمعیت ، استخراج جستجویی را پیاده سازی کرده است. روی بهترین تعداد خوشه ها تمرکز میکند و سپس به تدریج به طرف یافتن مراکز حرکت می کند. علاوه بر آن، استراتژی خوشه بندی حداکثر محدوده

صفات در جمعیت اولیه بکار رفته است که این به منظور غلبه بر حساسیت الگوریتم خوشه بندی برای تقسیم بندی اولیه انجام گرفته است. این الگوریتم برای شروع جمعیت اولیه طی آزمایشاتی، روش نمونه تصادفی را انتخاب کرده است. زیرا در استراتژی بکار رفته آن این روش نمونه تصادفی مراکز خوبی برای داده ها مشخص کرده است. برای تابع ارزیابی آن از مقدار همبستگی داخلی نسبت به خارجی خوشه انجام داده است. عملیات ژنتیکی: ۱- طی دو مرحله انتخاب تلاش می کند که نواحی امید بخش آن را انتخاب نماید. $f(X_i)$ تابع ارزیابی می باشد. N_s تعداد کروموزوم در جمعیتی که در همان تابع $f(X_i)$ است و S_i مجموعه ترکیبی از کروموزوم N_s است. عملیات به صورت معیار (۷-۲) انجام می گیرد.

$$P_s(i) = \left(N_s + \frac{\sum_{x_j \in f^a(X_j)} f^a(X_j)}{f^a(X_i)} \right)^{-1} = \left(N_s + \sum_{x_1 \in S_i} \left[\frac{f(X_j)}{f(X_i)} \right]^a \right)^{-1} \quad (7-2)$$

و مرحله تقاطع نیز از روش تقاطع موازی در جمعیت های مختلف بکار گرفته شده است. و جهش دو مرحله ای نیز با معیار احتمالات دوم بدست می آید. که در اینجا k_{con} مراکز خوشه است. عملیات به صورت معیار (۸-۲) انجام می گیرد.

(۸-۲)

$$p_m(i) = \begin{cases} 0.1 & \text{if } f \\ e^{-f} & \text{otherwise} \end{cases}$$

و این عملیات را تا زمانیکه بیش از تعداد تکرار مشخص شده باشد پایان می دهد. [۴۹]

۲-۳ پیشینه جنبه کاربردی مورد مطالعه

۲-۳-۱ چارچوب تئوریک و فرضیات عمده در رابطه با وفاداری مشتریان

وفاداری مشتری مفهومی است که بطور وسیع در زمینه رفتار مشتریان برای سالیان زیاد مورد استفاده قرار می گیرد. بعنوان مثال دیک و باسو (۱۹۹۴) به وفاداری مشتری بعنوان قدرت برقراری رابطه بین نگرش فرد در مورد موجودیت^{۳۶} (مارک، خدمت، انبار، فروشنده و...) و تکرار خرید توجه نموده اند.

بطور کلی سه مفهوم برای وفاداری مشتریان مشخص شده است:

۱- وفاداری بعنوان یک نگرش ابتدایی که گاهی اوقات منجر به برقراری یک رابطه با مارک مورد نظر می گردد.

۲- وفاداری عمدتاً براساس رفتار آشکار شده^{۳۷} (مانند الگوی خریدهای گذشته)

۳- خرید معتدل شده بوسیله نگرش افراد. [۵۰]

۱- وفاداری بعنوان یک نگرش ابتدایی

محققان معتقدند آنجا که یک تعهد فکری قوی نسبت به یک مارک وجود دارد، وفاداری نیز دیده می شود. این نگرش که ممکن است که بوسیله پرسش از افراد در مورد اینکه به چه میزان آنها به یک مارک علاقه دارند، و نسبت به آن احساس تعهد میکنند، آنها را به دیگران توصیه می کنند، و در رابطه با آن در مقایسه با سایر مارک ها، اعتقادات و احساسات مثبت دارند، اندازه گیری شود. قدرت این نگرشها، پیش بینی کننده مهمی برای خرید یک مارک می باشد. این آن چیزی است که اولیور (۱۹۹۷) وقتی که وفاداری مشتری را تعریف می کند، در ذهن دارد: "یک تعهد عمیق برای خرید مجدد یا مشتری شدن مجدد برای یک محصول یا خدمت در آینده، بوسیله تکرار خرید از مارک مشابه یا مجموعه ای از مارک های مشابه، علی الرغم تأثیرات موقعیتی و تلاشهای بازاریابی برای تأثیرگذاری بر روی تغییر رفتار". در زمینه های مربوط به تبلیغات جاری و تحقیقات مارک، این مدل برخی پشتیبانی ها را دریافت نموده است. به رغم توجه روانشناسی و

³⁶ Entity

³⁷ Revealed behaviour

جامعه شناسی به تفکرات مشتق شده از رویکرد های رابطه ای و رفتاری برای فهم وفاداری مشتریان، مفاهیم مربوط به وفاداری مشتریان بدون ضرر نمی باشند. تحقیقات تجربی سیستماتیک کمی برای تایید یا رد کردن این دیدگاه از وفاداری مشتریان وجود دارد. [۵۰]

۲- وفاداری براساس رفتار آشکار شده

این مفهوم، بسیار بحث انگیز می باشد. تحقیقات صورت گرفته مقدار زیادی داده در زمینه الگوهای خری در طول سالیان دراز در سراسر طبقه های تولید و برای بسیاری از کشورهای متنوع ایجاد نموده است. براساس تحقیقات صورت گرفته مشخص شده است که مشتریان بسیار کمی وجود دارند که نسبت به محصول ۱۰۰٪ وفادار هستند و یا کاملاً بیقید می باشند. بیشتر افراد در بین این دو سطح قرار می گیرند. (بعنوان مثال وفاداری به یک ترکیبی از مارک ها در یک طبقه محصول) از این دیدگاه، وفاداری بعنوان یک تمایل در حال رشد نسبت به خرید یک مارک از چندین مارک تعریف می شود. این محققان به تمرکز بر بازار بجای تمرکز بر روی فرد توجه دارند. معیارهای اصلی عملکرد، مقدار خرید و تعداد دفعات خرید و خریدهای تکراری برای یک دوره تعریف شده می باشد. [۵۰]

۳- خرید بوسیله نگرش افراد:

انسجام دادن به طرزفکر خریداران براساس یک تعریف از وفاداری به دو دلیل می باشد:

- ۱- تمایز بین وفاداری نگرشی و وفاداری غیرنگرشی در میان مشتریان مفید خواهد بود، زیرا مشخص می کند که چه کسانی وجه تعداد از مشتریان آسیب پذیر هستند، به همین دلیل، یک معیار در مورد اینکه چقدر طول می کشد تا مشتریان وفادار باقی بمانند، ارائه شده است.
- ۲- یک تعریف رفتاری صرف از وفاداری، که بتواند علل رفتار وفادارانه را توضیح دهد، وجود ندارد. [۵۰]

دسته بندی مشتریان به گروه های مختلف برای تحلیل آنها مفید خواهد بود. فرآیند داده کاوی، تعداد زیاد مشتریان را به گونه های مختلف دسته بندی می کند که مشتریان در هر دسته ویژگی های شبیه به هم دارند در حالی که ویژگی های مشتریان در هر گروه با مشتریان گروه های دیگر متفاوت است. در نتیجه می توانیم سلاقی مشتریان مختلف را بفهمیم تا بتوانیم رضایتمندی آنها را در

مورد محصولات به دست آوریم. به عنوان مثال در سرویس های مالی بلند مدت ، بانک اطلاعات زیادی را در مورد مشتریان جمع آوری می کند که شامل درآمد ، نوع و قالب زندگی و سابقه خدماتی است که به آنها ارائه شده است. در نتیجه خیلی از مشتریان ممکن است در خدماتی مانند وام گرفتن ، سپرده گذاری و خدمات مالی دیگر مثل هم باشند و گروه هایی تشکیل دهند که خصوصیات مشترکی دارند. با دسته بندی کردن مشتریان اشتراک هایشان را کشف می کنیم. و این باعث می شود که از یک طرف آنها را در سرمایه گذاری هایشان راهنمایی کنیم و از طرف دیگر سطح خدمات بانک را بالا ببریم که این خود باعث سوددهی بالای بانک خواهد شد [۵۱].

۲-۳-۲ تحلیل ارزش مشتریان RFM

ارزش طول عمر مشتریان مفاهیمی که در مدیریت روابط مشتری به وجود آمده است. پیپر^{۳۸} معتقد است که هدف CRM برای ایجاد نزدیک و روابط عمیق تر با مشتریان است تا طول عمر مشتریان را در سازمان به حداکثر می رسانند. یکی از قدرتمند ترین و ساده ترین مدل پیاده سازی CRM ممکن است. مدل RFM - تازگی^{۳۹}، توالی^{۴۰} و پولی^{۴۱} باشد. بولت^{۴۲} و وانسیک^{۴۳} RFM را به این صورت تعریف کرده اند: ۱- تازگی، پریود زمانی آخرین خرید: مقدار کمتری از آن به معنای احتمال بالاتری از تکرار خرید مشتریان ۲-توالی، تعداد خرید انجام گرفته شده در یک دوره خاص: که مقدار بالای آن به وفاداری مشتریان اشاره می کند. ۳- پولی، پول صرف شده در مدت زمان معین: که مقدار بالای آن نشان می دهد که شرکت ها بایستی بر روی این مشتریان تمرکز کنند. در برخی مقالات نیز بیشتر نویسندگان RFM وزن دار را پیشنهاد داده اند. [۵۲]

38 Pepers
39 Recency
40 Frequency
41 Monetary
42 Bult
43 Wansbeck

از این رو استراتژی های مختلفی را از نقطه نظر چگونگی تقسیم بندی ویژگی های رفتاری و درونی مشتریان را با مدل های مختلف مورد بررسی قرار داده ایم.

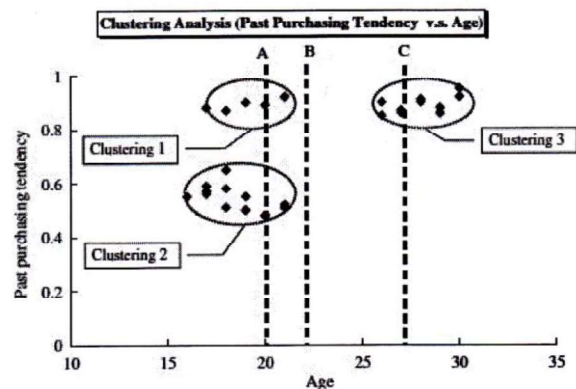
در مقاله ی با استراتژی AHP و کاوش داده ها با پیشنهاد محصولات بر مبنای ارزش طول عمر مشتری انجام داده است که در آن ابتدا بر روی داده ها عملیات پیش پردازش انجام می شود و پس از آن مقدار ارزش RFM استخراج می شود. با خوشه بندی kmean بر روی مقدار ارزش RFM که در نهایت خوشه های محصولات پیشنهادی و خوشه های مقدار ارزش پیشنهادی مشخص می شود. استراتژی AHP در سه مرحله انجام گرفته است. ۱- مقایسات زوجی نسبت متغیر های RFM که در آن مثلاً قوی به ضعیف مشخص شده است. ۲- ارزیابی سازگاری زوج ها که در آن با استفاده از شاخص سازگاری محاسبه می شود، که اگر شاخص سازگاری کمتر از ۰.۱ باشد. وزن دهی RFM مورد پذیرش قرار می گرفت. اگر شاخص سازگاری بیش از ۰.۱ باشد در وزن دهی RFM تجدید نظر می شود. ۳- مقایسه وزن های نسبی که در وزن دهی RFM کمک می کند. [۵۳]

کارهای زیادی در مورد مشتری مداری بانک آنلاین انجام گرفته شده است. رشد بی نظیر رقابت در تکنولوژی بانک بر خواسته از اهمیت نگهداری مشتریان دائم و بدست آوردن مشتریان جدید از طریق تحلیل رفتار مشتریان بر اساس پایگاه داده بانک انجام میگیرد. تعدادی از کارها در تحلیل پایگاه داده بانک از نقطه دیدگاه تحلیل رفتار مشتریان بررسی شده است. در مقالات برگزیده یک طرح دو مرحله ای از تحلیل رفتار مشتریان که از الگوریتم k-means و استنتاج قوانین وابستگی استفاده شده است. الگوریتم k-means برای شناسایی گروهی از مشتریان بر مبنای (تاخیر، ورود پی در پی، پولی) امتیاز رفتاری را در نظر میگیرد. و همچنین مشتریان را به گروه سود آور اصلی مشتریان تبدیل می کند، که با استفاده از استنتاج قوانین وابستگی گروهی از مشتریان را بوسیله ایجاد پروفایل مشتری مشخص کرده است. در این مقاله شناسایی مشتریان به وسیله مدل تحلیل رفتار مشتریان به منظور تشخیص مشتریان و انعطاف توسعه استراتژی بازار که دارای اثربخشی زیادی داشته است استفاده شده است. [۵]. و در مقالات دیگر نیز با عنوان پیش

بینی رفتار مشتریان در خوشه بندی و کشف قوانین وابستگی انجام گرفته است که سعی کرده اند، محصولات مورد علاقه مشتریان را ارائه داده شوند. از آنجایی که خرده فروشی بر روی خط ۴۴ امکان پیشنهاد محصولات گوناگون گسترده ای را می دهد، تمرکز روی محصولات ستاره ای شده است. بر طبق درجه ای وفاداری برای مشتریانی که محصول ویژه ای را می خرند، به صورت معیار (۴-۲) محاسبه می شود.

$$PPT_i = [Q/Q_0 \times (F + T/t) - \min] / (\max - \min) \quad (4-2)$$

۱- مقدار کل محصولات ویژه خریداری شده توسط مشتریان ۲- Q_0 مقدار آستانه . مشتری که مقدار کل آن بیش از مقدار آستانه باشد مشتری وفادار می باشد. نرخ مقدار کل به مقدار آستانه نشان می دهد که چه تعداد خرید کرده است. ۳- F اشاره به مشتری i دارد که چند بار محصول ستاره ای را خرید کرده است. F بالا در اغلب موارد خرید کرده است. ۴- T فریم زمانی بین اولین خرید و آخرین خرید می باشد. ۵- t دوره گردش کالا ستاره ای که به وسیله سیستم طراحی و تصمیم گیری می شود. اقلام با دوام T و اقلام مصرفی t نرخ T به t چه مدت طول کشیده که مشتری محصول ستاره ای را بخرد. در شکل (۴-۲) نمونه ای از دسته بندی نشان داده شده است. که در آن PPT و سن مشتریان خوشه بندی شده است.



شکل (۲-۴) نمونه خوشه بندی

با توجه به سرعت در حال بهبود تکنولوژی، در حال حاضر قادر به تشخیص تغییرات علاقه مندی مشتریان می باشد. با تحلیل بر روی صفحات وب، هنگامی که مشتری بر روی خط با کلیک کردن خرید خود را دنبال می کنند. و بر این فکر افتادند که داده های اخیر را از آن صفحه برای انعکاس دادن جریان علاقه مندی های محصول مشتریان استخراج کنند. همچنین تلاش کرده اند که بتوانند مشتریانی را که به احتمال زیاد محصولات ستاره ای را بوسیله مشتریان وفادار و بالقوه که براساس زمینه شخصی شان می باشد تقسیم بندی کنند [۵۴]. بر طبق سادرلند و کنول ۲۰۰۴ اساس مفاهیم اصل پارتو ۸۰ درصد تجارت درامدشان از ۲۰ درصد مشتریانشان است. اصل پارتو، بوسیله هافتر بیشتر توضیح داده شده است. کسی که مقاله اش را با عرصه سیاره های زنجیره ای انتشار داد. مدل پایرامید [۵۵] مشتریان را به پنج گروه خیلی ضعیف، کوچک، متوسط، بالا، خیلی بالا تقسیم بندی کرده است. که در [۵۶] از این مدل و با استفاده از خوشه بندی k-means عمل خوشه بندی را انجام داده است. تا مدیران عملیات تصمیم گیری در مورد مشتریان انجام دهند.

در مقاله ای دیگر رابطه بین محصول مبتنی بر وفادار و خوشه بندی بر پایه سری زمانی داده های سوپر مارکت انجام گرفته شده است. خوشه بندی روی ملاقات آنها انجام شده است. و کل هزینه هر هفته با استفاده از شبکه های عصبی و K-mean انجام می گیرد. نتایج خوشه بندی به صورت سری زمانی برای تحلیل کارایی سیستم امتیازدهی مشتریان ترسیم شده است. سیستم امتیاز دهی برای مشتریان وفادار در سوپرمارکت به منظور ارزیابی پیشنهاد شده است. که امتیاز مشتری بر پایه خرید محصولات کلیدی انجام می گیرد. این سیستم همیشه وفاداری را تشخیص نمی دهد. سیستم امتیازدهی بر اساس ارائه محصولات جدید تغییر پیدا می کند. [۵۷]

۲-۴ برگرفته از مطالعات

از جنبه ساختاری مسائل مورد بررسی بر روی انواع الگوریتم های خوشه بندی مشهور کار توسعه را انجام داده اند یکی از اینها الگوریتم ژنتیک بوده که با دید سراسری با موضوع برخورد می کند. در این الگوریتم با حالت های تصادفی که برخورد می کند، فضاهای مناسبی را برای جستجو ایجاد می کردند. اما غافل از آنکه استراتژیهای مورد مطالعه نرخ تعیین شده، تعداد اندازه جمعیت مناسب و تعداد خوشه ها تا حدودی کم رنگ دیده شده است. در تحقیق حاضر با حل این چالش سعی در بهینه کردن عملکرد خوشه بندی را داشته ایم. و از جنبه کاربردی نیز استراتژی های مختلفی برای دسته بندی مشتریان از نقطه نظر شناسایی رفتار انجام داده اند. تا بررسی شود که هر کدام از مشتریان چه نیازهایی دارند و بر اساس این نیازمندی ها، محصولات مورد علاقه آنها چه می باشد.

فصل سوم روش انجام کار

۱-۳ روش انجام کار

۱-۳-۱ مقدمه

الگوریتم ژنتیک به عنوان یک الگوریتم محاسباتی بهینه سازی، با در نظر گرفتن مجموعه ای از نقاط فضای جواب در هر تکرار محاسباتی، بنحو موثری نواحی مختلف فضای جواب را جستجو می کند. در مکانیزم جستجو گرچه مقدار تابع هدف تمام فضای جواب محاسبه نمی شود ولی مقدار محاسبه شده تابع هدف برای هر نقطه، در متوسط گیری آماری تابع هدف در کلیه زیر فضاهایی که آن نقطه بدان ها وابسته بوده، دخالت داده می شود و این زیر فضاها بطور موازی از نظر تابع هدف متوسط گیری آماری می شود. مشکلات k-means، از جمله تعداد خوشه بندی و همچنین گیر کردن در دام بهینه محلی، خوشه بندی k-means را به چالش کشانده است، ایده حاضر، الگوریتمی ارائه می شود در راستای حل این مشکلات، که با استفاده از خوشه بندی بر پایه ژنتیک بهینه که به طور خودکار تعداد خوشه را انتخاب و همچنین با حرکت موازی سعی در رفع گیر کردن بهینه محلی را انجام می دهد.

در شکل (۱-۳) مراحل کلی، فرآیند پیشنهادی نشان داده شده است. همانطور که مشاهده می کنید. گام اول مربوط به الگوریتم خوشه بندی پیشنهادی می باشد. که این گام با پیشنهاد دو دید از مساله پرداخته شده است. یکی در سطح ساختار با مطالعات بر روی انواع پراکندگی داده ها توانسته ایم با یکی از معیار هایی از جمله انحراف معیار، استراتژی دقت در خوشه بندی را با کمک ژنتیک انجام دهیم. که این کار باعث شده است. چالش های پیش رو را از جمله خودکار بودن تعداد خوشه بندی ها و دیگری گیر کردن در دام محلی الگوریتم، نسبت به دیگر خوشه بندی ها انجام گرفته شده است، این الگوریتم با نرم افزار متلب انجام گرفته شده است. در دید کاربردی نیز مساله را به

همراه ارائه مدلی در رفتار مشتری توانسته ایم مشتریان را با حالات درونی و بیرونی، تحلیل و فرمول بندی نماییم. و در نهایت این دو دید را با هم ترکیب کردیم، تا یک خوشه بندی با دقت و کارا را داشته باشیم. گام دوم مربوط به کشف قوانین بین محصولات مشتریانی که به عنوان مشتری وفادار شناخته شده اند بدست می آوریم که این کار با نرم افزار orange انجام داده ایم. و در گام آخر نیز محصولات مورد علاقه مشتریان وفادار را با استفاده از مقدار قوانین بدست آمده، برای مشتریان دیگر ارائه می دهیم.



شکل (۱-۳) مراحل روش انجام کار در یک نگاه

۳-۲-۲ بررسی گام اول از دید ساختاری

۳-۲-۱-۲ الگوریتم خوشه بندی پیشنهادی

در زیر شبه کد الگوریتم نشان داده شده است، که در آن از خط دو تا سیزده مربوط به انحراف معیار کلی مقادیر صفات می باشد که پس از آن با یک شرط، تعداد حالت های فضای جستجوی انتخاب می شود. در خط چهارده تعداد جمعیت نسل ها مشخص می شود و در خط پانزده تابع ارزیابی مشخص شده است. و از شانزده تا نوزده مربوط به عملگرهای ژنتیکی تقاطع، جهش و انتخاب می باشد که در این الگوریتم با آزمایشات مختلف نرخ های مورد نظر تعیین شده است. (در پیوست دو این تحقیق برنامه اصلی همراه با توضیحات آمده است)

```
1-begin
2-LOCATION=(xlsread('Clust3','Location'));
3-for i=1:size(LOCATION,1)
4-x(i,1)=var([LOCATION(i,2) LOCATION(i,3) LOCATION(i,4)] )
5-end
6-x=var(x,1)
7-v=sqrt(x)
8-if (v<rate)
9-num=2
10-else
11-num=3
12-end
13-num=3*num;
14-PopSize=input('Population size:');
15-FitnessFcn = for i=1:size(LOCATION,1)
for j=1(size(x,2)/3)
distances(j)=sqrt((LOCATION(I,2)-x(j))^2+(LOCATION(I,3)-
x(size(x,2)/3+j))^2+(LOCATION(I,4)-x(2*size(x,2)/3+j))^2);
end
16-numberOfVariables = num;
17-LOCATION=(xlsread('Clust3','Location'));
18-Options=gaoptimset('PlotFcns',my_plot,'PopInitRange',[0;1],
'PopulationSize',PopSize,'SelectionFcn',@selectionroulette,
'CrossoverFcn',{@crossoverScattered},'MutationFcn',{@mutationadaptfe
sible,0.9});
```



```

19-[x,fval] = ga(FitnessFcn,numberOfVariables,Options);
20-end

```

۲-۲-۱-۳ یافتن تعداد خوشه

ابتدا برای بدست آوردن از واریانس مقادیر صفات هر یک از اشیا را بدست می آوریم و در نهایت از مقادیر واریانس هر یک از اشیا یک واریانس کلی می گیریم و حاصل را در رادیکال قرار می دهیم یعنی در واقع انحراف معیار کلی واریانس اشیا را بدست می آوریم. این کار باعث می شود موقعیت اشیا نسبت به فضا حدس زده شود. در واقع هرچه پراکندگی اشیا در فضا زیاد باشد. تعداد خوشه ها نسبت به انحراف معیار کلی محاسبه می شود. بعد از انتخاب تعداد خوشه حال بایستی از چه روش هایی ژنتیکی در خوشه بندی استفاده شود.

فرمول واریانس به صورت فرمول زیر می باشد.

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{x})^2}{(n-1)} \quad (1-3)$$

که در آن هر یک از X_i مقدار اشیا از مقدار میانگین کاسته می شود. و پس از آن با استفاده از فرمول انحراف معیار که در ۲-۳ آمده است. در اینجا X_i مقدار واریانس کلی هر یک از اشیا می باشد.

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{x})^2}{(n-1)}} \quad (2-3)$$

مقدار حاصله را طبق آزمایشات مکرر که انجام داده ایم می توان با نرخ ۰.۰۱ مقایسه کرد. مثلا اگر S مقداری بزرگتری از نرخ باشد خوشه بندی با سه متغیر انجام می گیرد و اگر کمتر از نرخ باشد با دو متغیر انجام می گیرد. نرخ های پیشنهادی در محدوده ۰.۵ می باشند. یعنی اگر انحراف معیار بین ۰.۰۱ و ۰.۵ قرار گیرد عمل خوشه بندی با دو متغیر انجام می شود و اگر بین ۰.۵ و ۱ باشد خوشه

بندی با سه متغیر انجام می گیرد. در واقع هر چه با این نسبت مقدار انحراف معیار بزرگتر شود به همان میزان تعداد خوشه مناسب تخمین زده می شود. تا مساله حرکت مراکز، به راحتی در فضای جستجو حل شود.

۳-۲-۱-۳ محاسبه تابع ارزیابی

در اینجا هدف هر یک از صفات شی را با متغیر مقایسه می شود. که در نهایت مقدار مینیمم را در مجموع تابع ارزیابی قرار می دهد. هر یک از اشیا بوسیله مقدار k متغیر با بردار k بعدی مشخص شده اند. که در اینجا هدف به حداقل رساندن متغیرهای درون خوشه می باشد. داریم $\{s_i; i=1, 2, \dots, N\}$ مجموعه ای از صفات شی به تعداد N می باشد. و همچنین شرط ما بر این است که فقط یک اشیا در درون خوشه می تواند قرار گیرد. تابع ارزیابی مسالمان به این صورت خواهد بود که فرض کنیم R تا شی داریم و هر یک سه صفت را در خود دارند. ابتدا با تولید کروموزوم و تقسیم کروموزوم به سه قسمت (با استفاده از نرخ واریانس) که هر قسمت آن با صفات اشیا تفریق و به توان دومی رسانند بعد درکل اینها جمع می شوند و جذر آنها محاسبه می شود (فاصله اقلیدوسی). و این کار برای هر شی انجام می گیرد، در نهایت مقدار تابع ارزیابی بدست می آید. تابع ارزیابی در زیر نشان داده شده است.

$$f = \sum_{r=1}^R \min \sum_{i=1}^N \sum_{j=1}^M \sqrt{(S_i - X_j)^2 + \left(\frac{en_s}{\bar{s}} / \frac{en_x}{\bar{x}}\right)} \quad (3-3)$$

در اینجا S صفت مورد نظر، X متغیر r اشیا، N تعداد صفت، M تعداد متغیر، R تعداد اشیا می باشد. اختلاف نسبی ضریب تغییرات یک شی را با ضریب تغییرات کروموزوم تولیدی محاسبه می کنیم. به این معنا که هر چه قدر اختلاف کمتری صورت بگیرد مراکز خوشه از نظر حوزه انتخابی نقاط داده ای را تحت پوشش خود قرار می دهند. en_s انحراف معیار هر یک از اشیا می باشد و $\bar{s}$ میانگین هر یک از اشیا می باشد. en_x انحراف معیار متغیر تصادفی تولیدی می باشد. $\bar{x}$ میانگین مقادیر تصادفی تولیدی می باشد.

۳-۲-۴ اعمال مراحل ژنتیکی بر روی مجموعه داده و اعتبار سنجی

در این قسمت به منظور تخمین نرخ پیشنهادی در روش های ژنتیکی، مراحل ژنتیکی بر روی مجموعه داده ها iris.dat به دست آمده است. این مجموعه داده شده توسط گیاه شناس اندرسون، جمع آوری شده و شامل نمونه های تصادفی از گل، متعلق به سه گونه versicolor، setosa و virginica است. برای هر یک از گونه ها، ۵۰ مشاهدات طول کاسبرگ، عرض کاسبرگ، طول برگ، عرض برگ ثبت شده است. این کار بر روی مجموعه داده های iris جمع آوری شده است. در پیوست یک نمونه داده آمده است [۶۳]. به منظور ارزیابی و مقایسات مراحل ژنتیکی از شاخص دون برای اعتبار سنجی خوشه بندی بکار می رود. روش های ارزیابی خوشه های حاصل از خوشه بندی را به سه دسته تقسیم می کنند که عبارتند از:

- معیارهای بیرونی (External Criteria)
- معیارهای درونی (Internal Criteria)
- معیارهای نسبی (Relative Criteria)

هم معیارهای بیرونی و هم معیارهای درونی بر مبنای روش های آماری عمل می کنند و پیچیدگی محاسباتی بالایی را نیز دارا هستند. معیارهای خروجی عمل ارزیابی خوشه ها را با استفاده از بیش خاص کاربران انجام می دهند. معیارهای درونی عمل ارزیابی خوشه ها را با استفاده از مقادیری که از خوشه ها و نمای آنها محاسبه می شود، انجام می دهند.

پایه معیارهای نسبی، مقایسه بین شماهای خوشه بندی (الگوریتم بعلاوه پارامترهای آن) مختلف است. یک و یا چندین روش مختلف خوشه بندی چندین بار با پارامترهای مختلف روی یک مجموعه داده اجرا می شوند و بهترین شما خوشه بندی از بین تمام شماها انتخاب می شود. در این روش مبنای مقایسه، شاخص های اعتبارسنجی (Validity-Index) هستند. شاخص های ارزیابی بسیار متنوعی پیشنهاد شده اند که در این قسمت سعی می شوند تعدادی از رایج ترین آنها معرفی شوند.

شاخص‌های اعتبارسنجی برای سنجش میزان صحت (Goodness) نتایج خوشه‌بندی به منظور مقایسه بین روشهای خوشه‌بندی مختلف یا مقایسه نتایج حاصل از یک روش با پارامترهای مختلف مورد استفاده قرار می‌گیرند. در جدول زیر مجموعه‌ای از علائم استفاده شده در ادامه این بخش ارائه شده است. [۵۹]

جدول (۱-۳) مجموعه علائم بکار رفته در این بخش

علامت	مفهوم
n_c	تعداد خوشه‌ها
d	تعداد ابعاد
$d(x, y)$	فاصله بین دو داده
$\bar{X}_j$	امید ریاضی ژامین بعد داده‌ها
$\ X\ $	$\sqrt{X^T X}$ یک بردار ستونی است.
n_{ij}	تعداد داده‌های درون ژامین بعد از خوشه ژام
n_j	تعداد عناصر ژامین بعد از تمام داده‌ها
v_i	نقطه مرکز خوشه ژام
c_i	ژامین خوشه
$\ c_i\ $	تعداد داده‌های درون ژامین خوشه

۱- شاخص دون (Dunn Index)

این معیار توسط رابطه زیر تعریف می‌شود:

$$D = \min_{i=1 \dots n_c} \left\{ \min_{j=i+1 \dots n_c} \left(\frac{d(c_i, c_j)}{\max_{k=1 \dots n_c} (\text{diam}(c_k))} \right) \right\} \quad (4-3)$$

که $d(x, y)$ و $\text{diam}(c_i)$ در آن به ترتیب با روابط ۵-۳ و ۶-۳ محاسبه می‌شوند.

$$d(c_i, c_j) = \min_{x \in c_i, y \in c_j} \{d(x, y)\} \quad (5-3)$$

$$\text{diam}(c_i) = \max_{x, y \in c_i} \{d(x, y)\} \quad (6-3)$$

اگر مجموعه داده‌ای، دارای خوشه‌هایی جداپذیر باشد، انتظار می‌رود فاصله بین خوشه‌ها زیاد و قطر خوشه‌های (Diameter) آن کوچک باشد. در نتیجه مقداری بزرگ‌تر برای رابطه این معیار مقداری مطلوب‌تر است. معایب این معیار عبارتند از:

- محاسبه زمانبر
- حساسیت به نویز (قطر خوشه‌ها در صورت وجود یک داده نویزی می‌تواند بسیار تغییر کند).

۲- شاخص دیویس بولدین (Index Davies Bouldin)

این معیار از شباهت بین دو خوشه (R_{ij}) استفاده می‌کند که بر اساس پراکندگی یک خوشه (S_i) و عدم شباهت بین دو خوشه (d_{ij}) تعریف می‌شود. شباهت بین دو خوشه را می‌توان به صورتهای مختلفی تعریف کرد ولی بایستی شرایط زیر را دارا باشد.

- $R_{ij} \geq 0$
- $R_{ij} = R_{ji}$
- اگر S_i و S_j هر دو برابر صفر باشند آنگاه R_{ij} نیز برابر صفر باشد.
- اگر $s_j > s_k$ و $d_{ij} = d_{ik}$ آنگاه $R_{ij} > R_{ik}$
- اگر $s_j = s_k$ و $d_{ij} < d_{ik}$ آنگاه $R_{ij} > R_{ik}$

معمولا شباهت بین دو خوشه به صورت زیر تعریف می‌شود:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \quad (۷-۳)$$

که در آن d_{ij} و s_i با روابط زیر محاسبه می‌شوند.

$$d_{ij} = d(v_i, v_j) \quad (۸-۳)$$

$$s_i = \frac{1}{\|c_i\|} \sum_{x \in c_i} d(x, v_i) \quad (۹-۳)$$

با توجه به مطالب بیان شده و تعریف شباهت بین دو خوشه شاخص دیویس بولدین به صورت زیر تعریف می‌شود.

$$DB = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i \quad (۱۰-۳)$$

که R_i در آن به صورت زیر محاسبه می‌شود.

$$R_i = \max_{j=1 \dots n_c, i \neq j} (R_{ij}), i=1 \dots n_c \quad (۱۱-۳)$$

این شاخص در واقع میانگین شباهت بین هر خوشه با شبیه‌ترین خوشه به آن را محاسبه می‌کند. می‌توان دریافت که هرچه مقدار این شاخص بیشتر باشد، خوشه‌های بهتری تولید شده است.

۳- شاخص‌های اعتبارسنجی ریشه میانگین مربع انحراف از معیار (RMSSDT) و ریشه R (RS):

هرچند این شاخص‌ها معمولاً در اعتبارسنجی الگوریتم‌های سلسله مراتبی مورد استفاده قرار می‌گیرند ولی قابلیت ارزیابی نتایج سایر تکنیک‌های خوشه‌بندی را نیز دارا می‌باشند. در شاخص اعتبارسنجی RMSSDT (root - mean- square standard deviation) از واریانس خوشه‌ها استفاده می‌شود که به شکل رسمی می‌توان از رابطه ۱۲-۳ برای محاسبه آن استفاده کرد.

$$RMSSDT = \sqrt{\frac{\sum_{j=1..d} \sum_{k=1}^{n_j} (x_k - \bar{x}_j)^2}{\sum_{j=1..d} (n_j - 1)}} \quad (12-3)$$

با توجه به رابطه بالا و این که این معیار میزان همگنی خوشه‌ها را اندازه می‌گیرد، می‌توان دریافت که هرچه مقدار آن کمتر باشد نشان دهنده خوشه‌بندی بهتر داده‌ها است.

شاخص اعتبارسنجی RS (R Square) که با استفاده از رابطه ۱۳-۳، ۱۴-۳ و ۱۵-۳ تعریف می‌شود، معیاری برای بیان عدم تشابه بین خوشه‌ها است. به این شاخص درجه همگنی بین گروهی نیز گفته می‌شود. مقادیر آن به بازه اعداد بین ۰ تا ۱ محدود می‌باشد. که ۰ نشان دهنده نبودن هیچ تفاوتی بین خوشه‌ها و ۱ نشان دهنده وجود تفاوتی قابل توجه بین خوشه‌ها است.

$$RS = \frac{SS_t - SS_w}{SS_t} \quad (13-3)$$

$$SS_t = \sum_{j=1}^d \sum_{k=1}^{n_j} (x_k - \bar{x}_j)^2 \quad (14-3)$$

$$SS_w = \sum_{j=1..d} \sum_{k=1}^{n_j} (x_k - \bar{x}_j)^2 \quad (15-3)$$

در مقایسه این شاخص‌ها آزمایشات بر روی داده‌های درست و نادرست صورت گرفته است. که نشان می‌دهد تنها شاخص دون مقادیر صحیحی را محاسبه کرده است. [۵۹]

۱- مرحله انتخاب

روش های مختلفی برای الگوریتم های ژنتیک وجود دارند که می توان برای انتخاب ژنوم ها از آنها استفاده کرد. اما روش های لیست شده در پایین از معمولترین روش ها هستند. مشکل دیگر، که آن را نارس می نامیم این است که اگر یک ژنوم که فاصله اش با سایر ژنوم های نسل اش زیاد باشد (خیلی بهتر از بقیه باشد) و خیلی زود دیده شود (ایجاد شود) ممکن است محدودیت ایجاد کند و راه حل را به سوی جواب بهینه محلی سوق دهد. این اتفاق معمولاً در جمعیت های کم اتفاق می افتد. روش هایی برای مقابله با این مشکل وجود دارد که در زیر نمونه ای از آن آمده است. [۵۸]

- انتخاب Elitist: مناسبترین عضو هر اجتماع انتخاب می شود.
- انتخاب Roulette: یک روش انتخاب است که در آن عنصری که برآزش (تناسب) بیشتری داشته باشد، انتخاب می شود.
- انتخاب Scaling: به موازات افزایش متوسط عدد برآزش جامعه، سنگینی انتخاب هم بیشتر می شود و جزئی تر. این روش وقتی کاربرد دارد که مجموعه دارای عناصری باشد که عدد برآزش بزرگی دارند و فقط تفاوت های کوچکی آنها را از هم تفکیک می کند.
- انتخاب Tournament: یک زیر مجموعه از صفات یک جامعه انتخاب می شوند و اعضای آن مجموعه با هم رقابت می کنند و سرانجام فقط یک صفت از هر زیر گروه برای تولید انتخاب می شوند.

در تحقیق حاضر از روش چرخ رولت استفاده می کنیم. این انتخاب اولین بار توسط هالند پیشنهاد شده است. یکی از مناسب ترین انتخاب های تصادفی بوده که ایده آن، احتمال انتخاب می باشد. احتمال انتخاب متناظر با هر کروموزم، بر اساس برآزندگی آن محاسبه شده که اگر f_k مقدار برآزندگی کروموزم k ام باشد، احتمال بقای متناظر با آن کروموزم عبارت است از:

$$P_k = f_k / \sum_{i=1}^n f_i \quad (۱۶-۳)$$

حال کروموزوم ها را بر اساس p_k مرتب کرده و q_k که همان مقادیر تجمعی p_k می باشد به صورت زیر بدست می آید

$$q_k = \sum_i^k P_i \quad (17-3)$$

چرخ رولت به این صورت کار می کند که برای هر کروموزوم یک عدد تصادفی بین یک و صفر تولید کرده و عدد مذکور بازایی که قرار گرفت ، کروموزوم متناظر با آن انتخاب می شود. قرار است در اینجا محدوده جستجوی داده ها، منطقه ای انتخاب می شود که کروموزوم مناسبی برای مراکز باشد.

۲- مرحله تقاطع

وقتی با روش های انتخاب کروموزوم ها انتخاب شدند، بایستی به طور تصادفی برای افزایش تناسبشان اصلاح شوند. روش های متفاوتی وجود دارند که از جمله تقاطع یک نقطه ، دونقطه ، پراکنده، تخمینی، ریاضی و متوسطه که در اینجا با آزمایشات انجام گرفته روش پراکنده انتخاب شده است. در روش پراکنده یک بردار دوبعدی که هر ژن آن صفر یا یک ، اگر یک باشد ژن انتخابی از والد اول و اگر صفر باشد ژن انتخابی از والد دوم انجام می گیرد. برای مثال اگر p_2, p_1 به صورت زیر باشد.

$$\begin{aligned} p1 &= [a \ b \ c \ d \ e \ f \ g \ h] \\ p2 &= [1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8] \end{aligned}$$

و مقدار دودویی به صورت [1 1 0 0 1 0 0 0] باشد. نتیجه فرزند به صورت زیر می باشد.

$$child1 = [a \ b \ 3 \ 4 \ e \ 6 \ 7 \ 8]$$

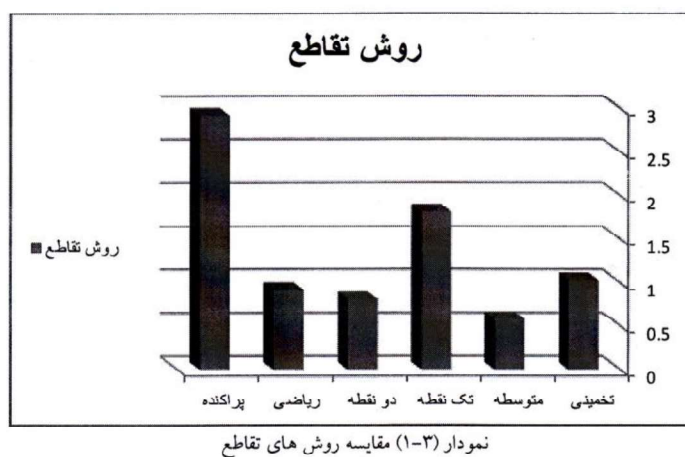
آزمایشات انجام شده بروی روش های متفاوت تقاطع ، با مجموع مقدار شاخص دون و میانگین مقدار شاخص دون در جدول (۲-۳) نشان داده شده است.

جدول (۲-۳) مقایسه روشها طی چهار بار آزمایش

مجموع مقدار شاخص دون	میانگین محاسبه شاخص دون	تعداد خوشه	عملیات تقاطع
۱،۰۳۰۵	۰،۴۱۲۲	۱۸	روش تخمینی
۲،۹۲۳۶	۱،۱۶۹۴۲	۱۶	روش پراکنده
۰،۵۷۹۳	۰،۲۳۱۷۲	۱۶	روش متوسطه

روش ریاضی	۱۴	۰,۳۶۶۸۸	۰,۹۱۷۲
روش تک نقطه	۱۴	۰,۷۲۷	۱,۸۱۷۵
روش دو نقطه	۱۶	۰,۳۲۴۰۴	۰,۸۱۰۱

و در نمودار (۱-۳) مشاهده می کنیم که در این روش ، روش پراکنده عملیات بهتری نسبت به بقیه دارد. این روش مقدار ارزیابی بالاتری نسبت به سایر روش ها دارد.



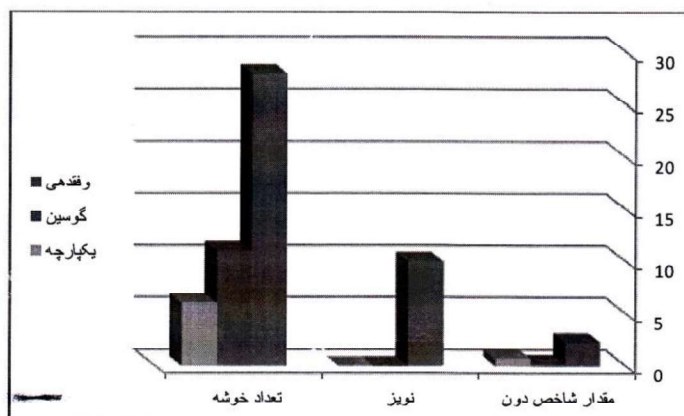
۳- مرحله جهش

در این مرحله روش های مختلفی از جمله جهش گاوسین، جهش یکپارچه و جهش تطبیق عملی وجود دارند . در این تحقیق به منظور غلبه بر مشکل بهینه محلی از جهش تطبیق عملی استفاده شده است. در جهش تطبیق عملی نسل ها به طور تصادفی در جهت تطبیق نسل ها موفق و ناموفق انجام می گیرد. که این باعث شده است مرزهای امیدوار کننده ای برای مراکز خوشه ایجاد

کند. در جدول (۳-۳) طی سه بار اجرا مقادیر بدست آمده را نشان می دهد. در نمودار (۲-۳) نیز می توان مشاهده کرد، در این نمودار تعداد خوشه ، مقدار شاخص دون و نویز (خوشه برای نویز) مشخص شده است که روش تطبیق عملی نسبت به دیگر روش ها عملکرد خوبی دارد. در نمودار (۳-۳) مقدار شاخص دون مشخص شده است.

جدول (۳-۳) مقایسه روش های جهش

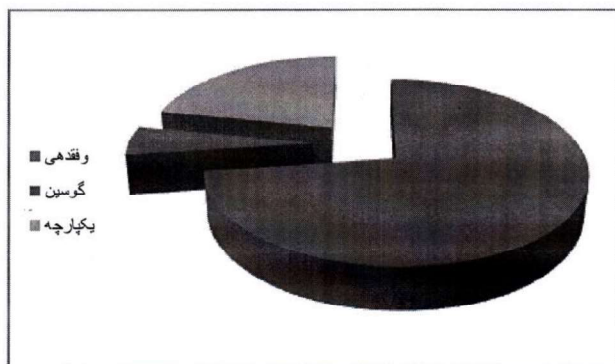
روش جهش	تعداد خوشه	مجموع مقدار شاخص دون	میانگین مقدار شاخص دون
روش گاوسین ^{۴۵}	۱۱	۰،۱۶۹۳۲	۰،۰۵۶۴۷۳
روش تطبیق عملی ^{۴۶}	۲۶	۲،۳۳۳۲	۰،۷۷۷۷۳۳
روش یکپارچه	۶	۰،۶۹۷۱	۰،۳۳۳۳۶۷



نمودار (۲-۳) مقایسه روش های جهش

^{۴۵}Gaussian

^{۴۶}adaptfeasible



نمودار (۳-۳) مقایسه روش های جهش با مقدار شاخص دون

بایستی برای این روش نرخ تعیین شود. که با آزمایشات انجام گرفته نرخ پیشنهادی با جمعیت ۱۰۰۰ و تقاطع پراکنده با نرخ ۰،۹ انجام گرفته شده است. جدول (۴-۳) مقدار آزمایشات نشان داده شده است. روش تطبیق عملی با استفاده از نسل های موفق و ناموفق قبلی مقدار ژن هایی مورد احتمال را تغییر می دهد که این کار با استفاده از شرایط مساوی و نامساوی انجام می دهد.

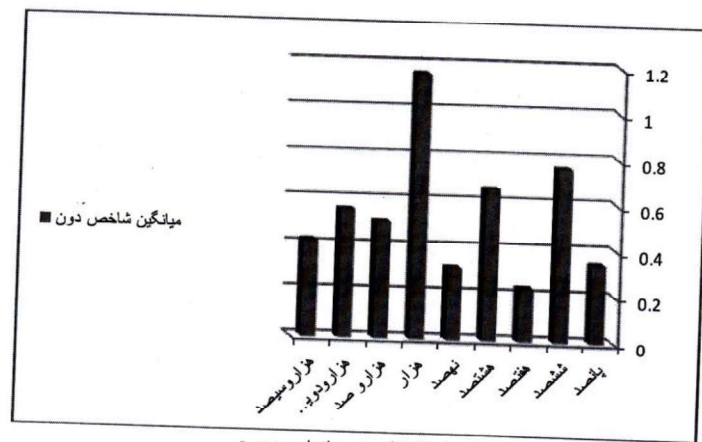
جدول (۴-۳) مقایسه نرخ های تعیین شده برای روش تطبیق عملی

مقدار شاخص دون	تعداد خوشه	نرخ
۰،۲۱۴۴	۹	۰،۱
۰،۶۹۷۵	۹	۰،۲
۰،۷۴۳۹	۱۴	۰،۳
۰،۱۱۰۷	۱۰	۰،۴
۱،۰۹۳۹	۱۰	۰،۵
۰،۱۱۵۱	۹	۰،۶
۰،۰۷۶۹	۹	۰،۷
۰،۸۱۴	۱۰	۰،۸
۱،۴۱۷	۹	۰،۹
۰،۰۳۴۸	۹	۱

از آنجائیکه مقدار انحراف معیار تعیین کننده در انتخاب تعداد خوشه می باشد. و هر چه تعداد خوشه ها بیشتر باشند. بایستی از نرخ نسبتاً بیشتری به تعداد قبلی برخوردار داشته باشند. چون کروموزم به تکه هایی تقسیم می شوند و این تکه ها تعریف حالتی از فضا می باشند. و نیاز است که انتخاب های بیشتری به منظور بهینگی فضای مساله انجام گیرد. حال فرض مقدار ۱۰۰۰ جمعیت را با دیگر مقادیر در محدوده آن آزمایش می کنیم و داده های بدست آمده در جدول (۳-۵) نشان داده شده است. در این جدول اندازه جمعیت بین حوزه ۵۰۰ و ۱۳۰۰ آزمایش شده است که مناسب ترین راه حل برای تقاطع تطبیق عملی و جهش یکپارچه با نرخ ۰.۹ در این جدول نشان داده شده است و همچنین نمودار (۳-۴) مقدار شاخص دون را نشان داده است.

جدول (۳-۵) انتخاب اندازه جمعیت

میانگین شاخص دون	مجموع شاخص دون	تعداد خوشه	اندازه جمعیت
۰.۳۳۹۸	۱,۰۱۹۴	۳۱	۵۰۰
۰.۷۵۶۷۶۶۶۷	۲,۲۷۰۳	۴۰	۶۰۰
۰.۳۳۰۵	۰,۶۹۱۵	۳۱	۷۰۰
۰,۶۶۱۰۶۶۶۶۷	۱,۹۸۳۲	۳۲	۸۰۰
۰.۳۱۱۵	۰,۹۳۴۵	۳۷	۹۰۰
۱.۱۵۶۱۳۳۳۳	۳,۴۶۸۲	۳۰	۱۰۰۰
۰.۵۰۷۰۳۳۳۳	۱,۵۲۱۱	۳۰	۱۱۰۰
۰.۵۵۹۱۳۳۳۳	۱,۶۷۷۴	۳۷	۱۲۰۰
۰.۴۲۱۳	۱,۲۶۳۹	۳۵	۱۳۰۰



نمودار (۳-۴) بهترین اندازه جمعیت

۲-۳ جنبه کاربردی پیشنهادی

۱-۲-۳ تعاریف

واژه دوره عمر مشتری به سطوح مختلف ارتباط بین مشتری و کسب و کار اشاره دارد. درک دوره عمر مشتری به این جهت که ارتباط مستقیمی با درآمد و مطلوبیت مد نظر مشتری دارد، مهم است.

معتقدند که سه روش برای افزایش ارزش مشتری وجود دارد:

۱- افزایش خرید مشتری از محصولی که قبلاً خریداری کرده است

۲- فروش محصولات با حاشیه سود بالاتر به مشتریان

۳- حفظ مشتری برای یک دوره زمانی طولانی تر

دوره عمر مشتری چارچوبی کلی برای درک رفتار مشتری بوده و شامل چهار گام اساسی زیر می باشد:

۱- (Prospects): افرادی که تاکنون جزء مشتریان نبوده ولی در بازار هدف قرار دارند.

۲- (Responder): افرادی که در بازار هدف بوده و علاقمندی به محصول یا خدمتی را نشان می دهند.

۳- (Active Customer): افرادی که به تازگی از محصول یا خدمتی استفاده کرده اند.

۴- (Former Customer): مشتریانی بد که صورتحساب خود را پرداخت نکرده، موجب ایجاد هزینه بالایی بوده و یا از محصولات رقیب خریداری می کنند. [۴]

دوره عمر مشتری چارچوبی مناسب برای بکارگیری داده کاوی جهت مدیریت روابط مشتری ارائه می دهد. این چرخه از جنبه درونی بیان می کند که چه اطلاعاتی در دسترس بوده و از جنبه بیرونی عنوان می کند که احتمالاً چه چیزی برای مشتری جذابیت دارد.

در این جا دو مورد مطرح است: اولی انبار داده های اکتشافی است که درباره Click Stream صنایع مختلف وجود دارد و شامل پیگیری رقابتهای مورد نظر می باشد. مورد دوم شامل اطلاعات رفتار اکتشافی است که در برخی از وب سایتها موجود است. داده کاوی می تواند سودآوری اکتشافات را در مواردی نظیر فعال شدن مشتریان و مدت زمان فعال بودن آنها و اینکه چگونه آنها محصول یا خدمتی را کنار می گذارند، پیش بینی می کند. در مجموع داده کاوی می تواند بیش از یک دوره زمانی جهت پیش بینی دقیق تغییرات مورد استفاده قرار گیرد. داده کاوی به سازمانها در شناخت الگوهای داده مشتری کمک می کند. برای مثال یک شرکت میتواند از داده کاوی جهت پیش بینی یک رویدادی خاص از دوره عمر مشتری (مانند بازنشستگی) استفاده نماید. در اینجا مشتریان را بر اساس مولفه های درونی و بیرونی تقسیم کرده ایم

۲-۳-۲ مولفه های بیرونی

در اینجا مشتریان بر اساس هفت مولفه از جمله زمان، تکرار مراجعه، قیمت خرید، درخواست، شکایت، تبادل نظر، بدهی تقسیم نموده ایم.

مولفه زمان: مربوط به میزان کل زمان سپری شده در آن سایت بوسیله ی مشتری

تکرار مراجعه: مربوط به کل مراجعات مشتری

قیمت خرید: مربوط به کل خرید نسبت به کل زمان

درخواست: ضریب امیدواری محصول را فرض بر تعداد درخواست آن محصول ضربدر ۴ و

در نهایت حاصل مجموع تمام ضریب امیدواری های کل محصول

شکایت: ضریب ناامیدی از گرفتن محصول فرض بر تعداد شکایت ضربدر منفی ۴ و در نهایت

حاصل مجموع تمام ضریب امیدواری کل محصول

تبادل نظر: ضریب میان امیدی محصول را فرض بر تعداد تبادل نظر آن محصول، ضربدر ۲ و

در نهایت حاصل مجموع تمام ضریب امیدواری کل محصول

بدهی: به ازای هر کالای بدهی ضربدر منفی ۴ و در نهایت حاصل مجموع تمام ضریب امیدواری

کل

محصول مشتریان بر مبنای click stream دسته بندی می شوند که به صورت ذیل می باشد:

اینجا مشتریان با مولفه های موجود احتمالات را گرفته و در خوشه بندی مناسب بر اساس مولفه

های درونی را انجام می دهیم.

احتمالات هریک از این ها با:

$$P=((T*M)/F+(S*-4)+(D*4)+(t*2)+(B*-4))$$

F: مربوط به کل مراجعات مشتری

M: مربوط به کل خرید نسبت به کل زمان

T: مربوط به میزان کل زمان سپری شده در آن سایت بوسیله ی مشتری

S: ضریب ناامیدی از گرفتن محصول فرض بر تعداد شکایت ضربدر منفی ۴ و در نهایت حاصل مجموع تمام ضریب امیدواری کل محصول

D: ضریب امیدواری محصول را فرض بر تعداد درخواست آن محصول ضربدر ۴ و در نهایت حاصل مجموع تمام ضریب امیدواری های کل محصول

t: ضریب میان امیدی محصول را فرض بر تعداد تبادل نظر آن محصول ضربدر ۲ و در نهایت حاصل مجموع تمام ضریب امیدواری کل محصول

B: به ازای هر کالای بدهی ضربدر منفی ۴ و در نهایت حاصل مجموع تمام ضریب امیدواری کل محصول

مشتریان بر مبنای click stream دسته بندی می شوند که به صورت ذیل می باشد. در اینجا فقط برای مشتریان (Prospects) T/F احتمال گرفته می شود. و برای مشتریان Active و Former به دلیل نزدیکتر شدن مقادیر به مشتریان prospect ، حاصل تقسیم بر عدد ۱۰۰۰۰۰۰۰ می شود. این کار باعث ارائه الگو مشتریان می باشد. و همچنین برای مشتریان Responder نیز تقسیم بر ۱۰۰ انجام گرفته است. این فرمول دهی نشان می دهد که داده ها به طور دقیق با مساله زیر ساختی خوشه برخورد می کنند تا مدیران در تعاملات با مشتریان چه از نظر دقت الگوریتم خوشه بندی و چه از نظر ویژگی های درونی و بیرونی مشتریان با آنها رابطه پایداری را برقرار کنند.

جدول (۳-۶) فرمول بندی برای دسته های مشتری

Former Customer	Active Customer	RESPONDER	Prospects
زمان	زمان	زمان	زمان
تکرار مراجعه	تکرار مراجعه	تکرار مراجعه	تکرار مراجعه
قیمت خرید	قیمت خرید	درخواست	*
بدهی	درخواست	تبادل نظر	*
درخواست	شکایت	*	*
شکایت	تبادل نظر	*	*
تبادل نظر	*	*	*

۳-۳-۳ مولفه های درونی

در اینجا مشتریان را بر اساس مشخصاتشان تقسیم نموده ایم که از جمله: سن و توان مالی و تحصیلات و ارزش محل زندگی، در نهایت این مولفه ها با مولفه های بیرونی خوشه بندی مناسب می شوند. پس از خوشه بندی بر روی محصولات مشتریان فعال کشف قوانین وابستگی انجام می دهیم تا قوانین به مشتریان خوشه بندی شده اعمال شود. در فصل بعدی به صورت مطالعه موردی فروشگاه لوازم خانگی ناریلا انجام گرفته شده است.

مقایسه و نتایج

۴-۱ مقدمه

در این فصل الگوریتم پیشنهادی را از نظر دقت حرکتی به منظور نمایش با الگوریتم های kmeans ، FCM مقایسه کرده ایم و همچنین با استفاده از شاخص های اعتبار سنجی خوشه با الگوریتم های iso data ، FCM و Gka بر روی پایگاه داده iris مورد مقایسه قرار داده ایم. و در پایان الگوریتم پیشنهادی را به صورت مطالعه موردی بر روی فروشگاه لوازم خانگی ناربلا انجام داده ایم.

۴-۲ مقایسه الگوریتم پیشنهادی از نظر دقت حرکتی در داده های خیلی کوچک

۴-۲-۱ نگاهی به جزئیات k-mean

در رابطه با الگوریتم پیشنهادی الگوریتم k-means استاندارد (خوشه بندی) که تمام الگوریتم های پیشنهادی با آن مقایسه می شوند. که این کار را با استفاده از کد نویسی متلب انجام داده ایم. لازم به ذکر است برای الگوریتم های مورد مقایسه در اینجا تعداد خوشه را از قبل مشخص شده می باشد .
قطعه برنامه زیر به شکل زیر است:

```
opts = statset('Display','final');
[idx,ctr] = kmeans(clust3,3,...
    'Distance','city',...
    'Replicates',5,...
    'Options',opts);
for i=1:14
    if (idx(I,i)==3)
        plot3(clustering(I,1),clustering(I,2),clustering(I,3),'sr','MarkerFaceColor','r','MarkerSize',10)
        hold on
    end
    if (idx(I,i)==2)
        plot3(clustering(I,1),clustering(I,2),clustering(I,3),'sb','MarkerFaceColor','b','MarkerSize',10)
    end
end
```

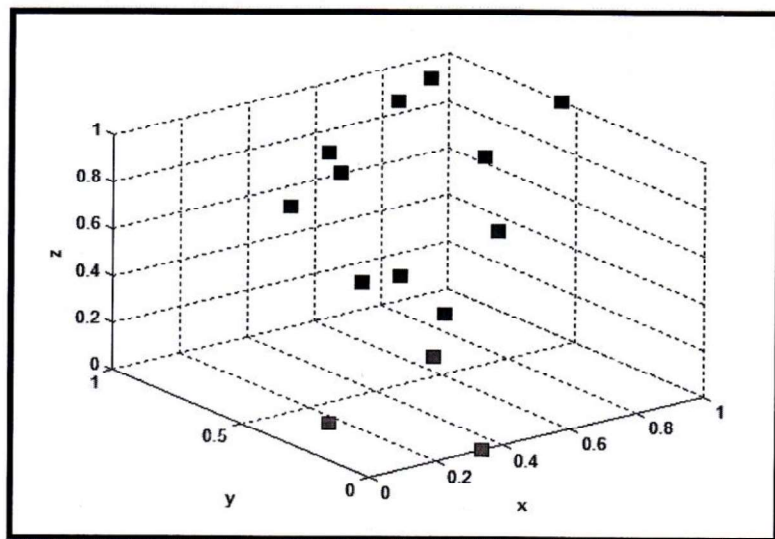


```

if (idx(I,1)==1)
plot3(clustering(I,1),clustering(I,2),clustering(I,3),'sr','MarkerFaceColor','g','MarkerSize',10)
end
hold on
end
hold on
xlabel('x');ylabel('y');zlabel('z');
grid on;
hold on;

```

در شکل (۱-۴) نتیجه الگوریتم k-means نشان داده شده است. این الگوریتم چیدمان متفاوتی با الگوریتم پیشنهادی دارا می باشد.



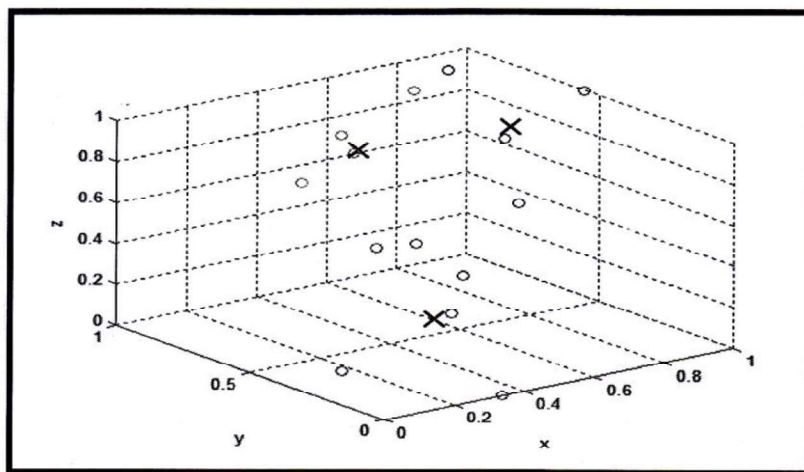
شکل (۱-۴) خوشه بندی k-means با نمایش سه بعدی

۲-۲-۴ نگاهی به جزئیات fuzzy mean

الگوریتم دیگر به نام خوشه بندی فازی fuzzy mean که داده ها به صورت میزان تعلق
سنجیده شده است. قطعه برنامه به صورت زیر می باشد.

```
[center,U,objFcn] = fcm(clusterfcm,3);
maxU = max(U);
index1 = find(U(1, :) == maxU);
index2 = find(U(2, :) == maxU);
index3 = find(U(3, :) == maxU);
plot3(clusterfcm(index1,1),clusterfcm(index1,2),clusterfcm(index1,3),'linestyle',...
'none','marker','o','color','g');
hold on
plot3(clusterfcm(index2,1),clusterfcm(index2,2),clusterfcm(index2,3),'linestyle',...
'none','marker','o','color','r');
hold on
plot3(clusterfcm(index3,1),clusterfcm(index3,2),clusterfcm(index3,3),'linestyle',...
'none','marker','o','color','b');
hold on
plot3(center(1,1),center(1,2),center(1,3),'kx','markersize',15,'LineWidth',2)
plot3(center(2,1),center(2,2),center(2,3),'kx','markersize',15,'LineWidth',2)
plot3(center(3,1),center(3,2),center(3,3),'kx','markersize',15,'LineWidth',2)
hold on
xlabel('x');ylabel('y');zlabel('z');
grid on
hold on;
```

در شکل (۲-۴) نتیجه الگوریتم نشان داده شده است. که به میزان تعلق آن داده ها را انتخاب کرده است.



شکل (۲-۴) خوشه بندی FCM

۲-۲-۴ نگاهی به جزئیات GA پیشنهادی

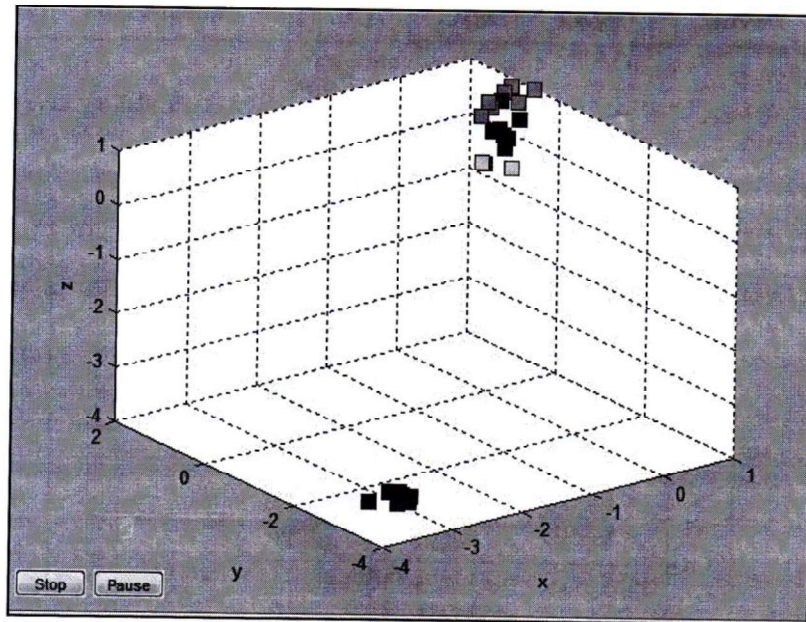
در اینجا فقط تابع ترسیم نقاط را آورده ایم، برنامه اصلی مثل همان کد نویسی پیوست دو می باشد. شکل ۳-۴ نقاط پیشنهادی مراکز به سمت داده ها حرکت می کنند و داده هایی که به مرکز خوشه نزدیکترند متعلق به همان خوشه می باشند. در این شکل به سه صورت خوشه بندی شده اند.

```
function state = Drawing(Options,state,flag,LOCATION,num)
[unused,i] = min(state.Score);
x=state.Population(i,:);
for i=1:size,LOCATION,1)
for j=1:(size(x,2)/3)
distances(j)=sqrt((LOCATION(i,2)-x(j))^2+(LOCATION(i,3)-
x(size(x,2)/3+j))^2+(LOCATION(i,4)-x(2*size(x,2)/3+j))^2);
end
distances
[min_distance,assign(i)]=min(distances)
end
```

```

for i=1:size(LOCATION,1)
plot3(LOCATION(i,2),LOCATION(i,3),LOCATION(i,4),'sr','MarkerFaceColor'
',[3*(assign(i))/num,3*(assign(i))/num,3*(assign(i))/num],'MarkerSize'
',10
xlabel('x');ylabel('y');zlabel('z');
grid on;
hold on;
end
plot3(x(1:size(x,2)/3),x((size(x,2)/3+1):2*size(x,2)/3),x(2*size(x,2)
/3+1:size(x,2)),'sr','MarkerFaceColor','r','MarkerSize',10);
hold off;

```



شکل (۴-۳) خوشه بندی ژنتیک پیشنهادی

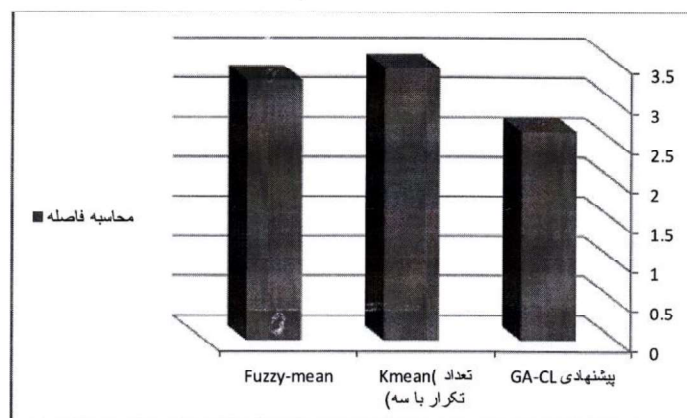
۳-۴ مقایسات مجموع فواصل درستی عملکرد خوشه

در اینجا ، با استفاده از رابطه فاصله مربع خطا نسبت به مرکز خوشه به منظور دقت حرکتی مرکز به خوشه ها مقایساتی را انجام دادیم .

جدول (۱-۴) مقایسه الگوریتم ها

الگوریتم	خوشه اول	خوشه دوم	خوشه سوم	جمع کل
پیشنهادی GA-CL	۰,۸۱۲۷	۰,۴۶۶۲	۱,۳۵۷	۲,۶۳۵۹
Kmean (تعداد تکرار با سه)	۱,۵۶۸۷	۱,۱۴۲۲	۰,۷۱۶۳	۳,۴۲۷۲
Fuzzy-mean	۰,۶۰۶۱	۰,۶۹۳۸	۱,۹۷۷۴	۳,۲۷۷۳

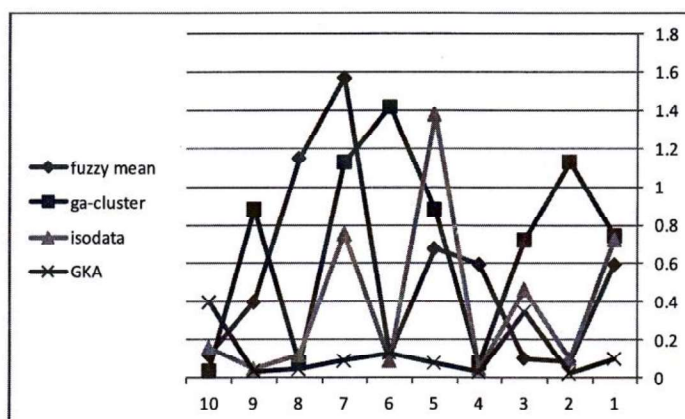
در نمودار (۱-۴) مقایسات نشان داده شده است ، که هر چه مقدار فاصله اقلیدوسی کمتری داشته باشیم . خوشه بندی مناسبی را در بر خواهیم داشت.



نمودار (۱-۴) محاسبه فاصله اقلیدوسی برای الگوریتم ها

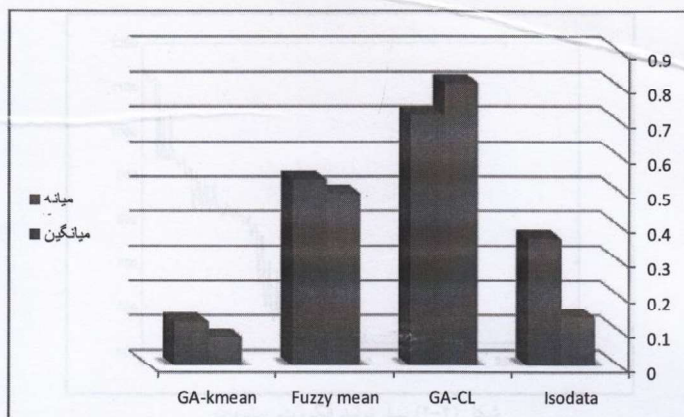
۴-۴ مقایسه الگوریتم با استفاده از شاخص دون

در این بخش الگوریتم پیشنهادی با الگوریتم Fuzzy mean ، iso data و GKA را مقایسه کرده ایم. هر یک از این الگوریتم ها بر روی مجموعه داده iris طی ده بار تکرار اجرا انجام داده ایم. در الگوریتم fuzzymean عملیات خوشه بندی با مقداردهی انجام گرفته شده است. ولی در روش isodata و GKA ژنتیک پیشنهادی به طور اتوماتیک صورت گرفته است. نمودار (۲-۴) مقدار اعتبار الگوریتم ها در هر بار اجرا نشان داده است.



نمودار (۲-۴) مقدار اعتبار الگوریتم ها در هر بار اجرا

نمودار (۳-۴) میانه و میانگین شاخص اعتباری برای هر الگوریتم نشان داده شده است.

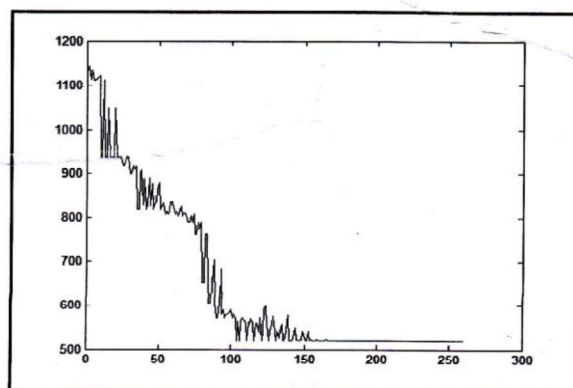


نمودار (۳-۴) میان و میانگین اعتبار خوشه بندی

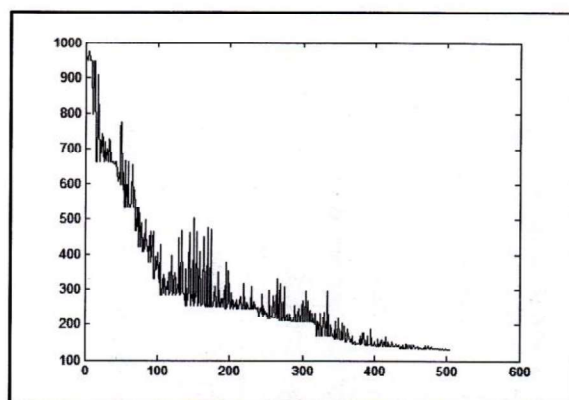
نتایج نشان می دهد که الگوریتم پیشنهادی در مقایسه با الگوریتم های نامبرده از میانگین و میان مقادیر بالاتری برخوردارند.

۵-۴ مقایسه نسل ها

در این بخش مقایساتی بر روی الگوریتم پیشنهادی و الگوریتم ژنتیک مبتنی بر kmean انجام گرفته شده است. شکل ۴-۴ الگوریتم پیشنهادی ، نسل های تولیدی را طی فرآیند حل مساله نشان داده شده است. و شکل ۵-۴ الگوریتم ژنتیک مبتنی بر فاصله اقلیدوسی نشان داده است. همانطور که مشاهده می کنیم در شکل ۴-۴ همانطور که مشاهده می کنید، نسبت تولید نسل ها از خیزش خوبی برخوردار است و حرکت نسل ها به سمت بهینه سراسری با استفاده از مراحل ژنتیکی از انعطاف بالایی برخوردار است.



شکل (۴-۴) نسل تولید الگوریتم پیشنهادی

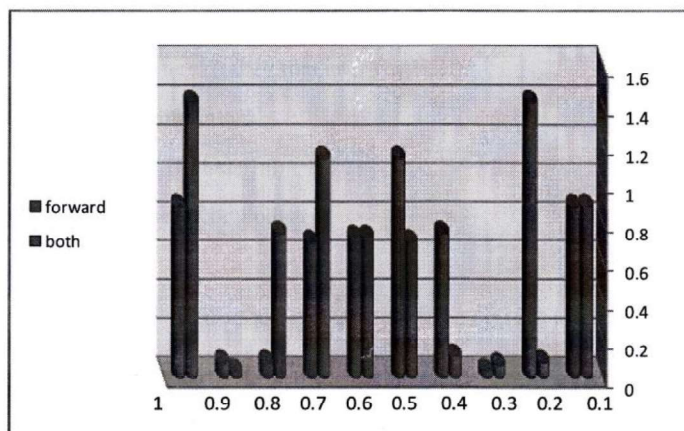


شکل (۵-۴) نسل تولید الگوریتم ژنتیک مبتنی بر kmean

۶-۴ مهاجرت نسل ها

در این بخش در راستای بهینه کردن الگوریتم بر روی مهاجرت نسل ها تمرکز کرده ایم . مهاجرت به این معنا که چگونه افراد از یک زیر جمعیت به زیر جمعیت دیگر حرکت می کنند به شکل عموم

دو روش برای این حرکت مهاجرت مشخص شده است. روش رو به جلو^{۴۷} مهاجرت از سمت زیر جمعیت گذشته انجام می گیرد. به این معنا که n امین زیر جمعیت به سمت $n+1$ زیر جمعیت عمل مهاجرت انجام می گیرد. روش دو تایی^{۴۸}، عمل مهاجرت از n امین زیر جمعیت به $n-1$ و $n+1$ زیر جمعیت انجام می گیرد. پس از انتخاب عمل مهاجرتی بایستی مشخص کنیم چه تعداد نسل از یک زیر جمعیت به زیر جمعیت دیگر مهاجرت می کنند. مثلاً نسل هایی از یک زیر جمعیت ۵۰ نسلی به یک زیر جمعیت ۱۰۰ نسلی قرار است عمل مهاجرت را انجام دهند. با فرض اینکه ۰/۱ نرخ مهاجرت می باشد که بر ۵۰ نسل زیر جمعیت اولیه ضرب می شود. در نتیجه حاصل آن ۵ نسل به زیر جمعیت دیگر انتقال می یابند. در نمودار ۴-۴ آزمایش با دو روش همراه با نرخ های مشخص شده انجام گرفته شده است.



نمودار (۴-۴) آزمایش مهاجرت ها با دو روش

⁴⁷ Forward

⁴⁸ Both

همانطور که در آزمایشات می بینیم ، هیچگونه رشد اعتبار شاخص دون را در پی نداشتیم. در نهایت عمل مهاجرت را به روشمان اضافه نمی کنیم.

۴-۲ مطالعه موردی جنبه کاربردی

در زیر شبه کدی از روند پیشنهادی نشان داده شده است.

Begin

1-input information inter-customer(salary,age)

2-input information outer-customer(T,M,F,S,D,tab,bed)

3-sum=((T*M)/F+(S*-4)+(D*4)+(tab*2)+(bed*-4))

4-function cluster GL-cluster(salary,sum,inter-customer)

5-pakage output function cluster

6-function distance-point(cluster[n])

7-function association-rule for product

End.

در اینجا نمونه ایی از شصت مشتریان فروشگاه لوازم خانگی ناریل را در محیط EXCEL بر اساس مولفه های بیرونی هر یک از مشتریان را وارد کرده و فرمول مربوطه که در فیلد SUM ذخیره شده اند .احتمال مشتریان را بدست می آوریم .جدول ۴-۲ نمونه داده مولفه های بیرونی را نشان می دهد.

جدول (۲-۴) نمونه داده مولفه های بیرونی

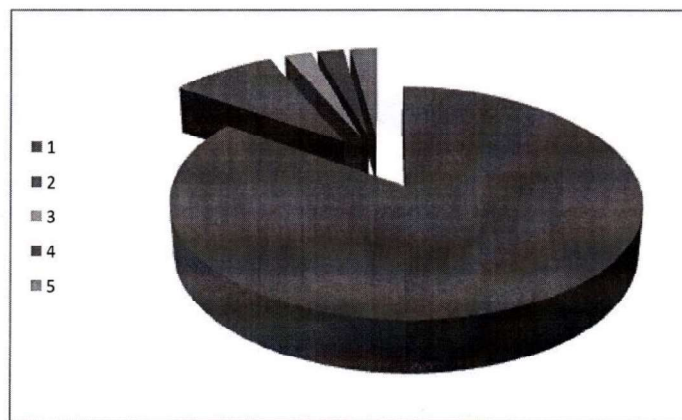
id	T	F	M	t	D	S	B sum	customer
1	3	10						0.3 prospect
2	2	15						0.13333333 prospect
3	3	27						0.11111111 prospect
4	5	23						0.217391304 prospect
5	5	14						0.357142857 prospect
6	7	12						0.583333333 prospect
7	5	18						0.277777778 prospect
8	6	5						1.2 prospect
9	98	55						1.781818182 prospect
10	2	8						0.25 prospect
11	13	3		1	8			0.34 responder
12	2	4		1	5			0.22 responder
13	38	17		9	6			0.42 responder
14	67	19		8	5			0.36 responder
15	29	43		3	3			0.18 responder
16	40	54	65000000	7	4	7		0.481481501 Active
17	43	34	54000000	6	3	6		0.682941176 Active
18	32	23	39000000	4	3	5		0.542608696 Active
19	33	41	43500000	5	7	4		0.350122171 Active
20	97	33	42000032	4	2	2		1.234546475 Active
21	87	36	67000000	5	6	23		0.161916087 Active
22	32	52	20980000	5	5	9		0.129107632 Active
23	5	17	8		6	7	3	-13.64705882 Former
24	34	19	6	4	6	6	4	2.736842105 Former
25	43	53	3	3	3	4	6	-19.58803774 Former

نمونه ای از مولفه های درونی مشتریان از جمله توان مالی مشتریان و سن مشتریان در جدول (۳-۴) آمده است.

جدول (۴-۲) توان مالی و سن مشتریان

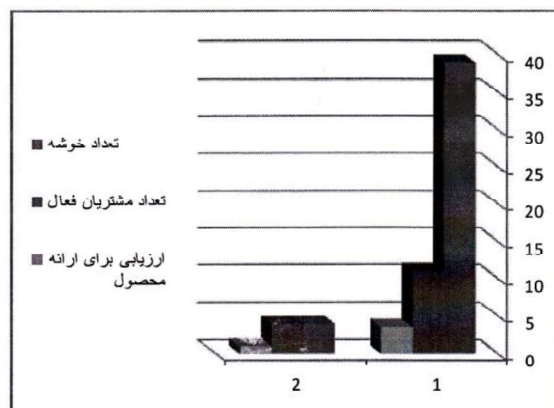
tavan mali	age
2000000	25
1200000	43
800000	34
500000	35
750000	35
900000	36
340000	35
800000	34
640000	38
760000	40
450000	47
590000	48
800000	50
730000	50
640000	37
550000	35
450000	34
430000	60
560000	45
640000	8
380000	44
490000	21
480000	23
580000	44

نمودار ۵-۴ عملیات خوشه بندی که از خوشه بندی ژنتیک پیشنهادی بدست آمده در زیر نشان داده شده است که حاصل خوشه بندی بر روی ۶۰ مشتری، پنج خوشه می باشد، که سه خوشه داده نويز را اختیار کرده است و دو خوشه را می توان برای آنها محصولات مورد علاقه مشتریان را ارائه داد.



نمودار (۵-۴) حاصل خوشه بندی

همانطور که در نمودار (۶-۴) مشاهده می کنید. تعداد مشتریان در هر خوشه، تعداد مشتریان فعال در هر خوشه و ارزیابی هر خوشه با محاسبه نسبت تعداد مشتریان به تعداد مشتریان فعال صورت گرفته است.



نمودار (۴-۶) ارزیابی هر خوشه به منظور ارائه محصول

تکامل تکنولوژی پایگاه داده و استفاده فراوان آن در کاربردهای مختلف سبب جمع آوری حجم فراوانی داده شده است. این داده های فراوان باعث ایجاد نیاز به ابزارهای قدرتمند برای تحلیل داده ها گشته، زیرا در حال حاضر به لحاظ داده ثروتمند هستیم ولی دچار کمبود اطلاعات میباشیم. کشف دانش در پایگاه داده فرایند شناسایی درست، ساده، و مفید الگوها و مدل‌های قابل فهم در داده ها می باشد.

کشف قوانین وابستگی^{۴۹} از مهمترین راههای تحلیل داده ها می باشد به کشف قوانین وابستگی بین کشف قوانین وابستگی داده های داخل یک پایگاه داده می پردازد. منظور از کشف وابستگی یافتن قوانین در موردی است که با هم اتفاق می افتند مثلاً اجناسی که در یک فروشگاه احتمال خرید همزمان $A \Rightarrow B$ آنها زیاد است. برای ارزیابی قوانین وابستگی از دو پارامتر استفاده می شود [۶۰] پشتیبانی درصد رکوردهای شامل آیتم های الف و ب با یکدیگر را نشان می دهد. از این معیار برای یافتن وابستگی میان آیتم های مختلف استفاده می شود. اطمینان: اطمینان درصد رکودهایی که هم آیتم الف و هم آیتم ب را شامل می شوند، محاسبه می کند. این معیار برای محاسبه درصد اطمینان به قوانین بدست آمده محاسبه می گردد. به عنوان مثال در پایگاه اطمینان^{۵۰} با نان $\Rightarrow$ کره داده ی یک فروشگاه، قانون اطمینان برابر با ۷۵٪ و قانون پشتیبان ۵۰٪ می باشد در اینجا ۵۰٪ به این

^{۴۹} Association rule

^{۵۰} Confidence

معناست که ۵٪ از کل تراکنش های پایگاه داده شامل هر دو بوده اند که نشان دهنده ی میزان کره و نان به وفور این دو عنصر در پایگاه داده پیدا شده است. همچنین ۷۵٪ از نان بوده اند، آیتم کره تراکنش هایی که شامل آیتم نان هم مشاهده شده است. در زیر معیار پشتیبان و اطمینان نشان داده شده است.

$$support = \frac{(XUY).count}{n} \quad (13-4)$$

$$confidence = \frac{(XUY).count}{X.count} \quad (14-4)$$

در داده های واقعی زمان یک پارامتر در ساعات اولیه ی کره و نان مهم است، به عنوان مثال اطمینان و پشتیبان در یک روز بین ۸ تا ۱۰ صبح دارای مقادیر بالایی می باشد، در حالی که در ساعات دیگر روز این دو مقدار بسیار ناچیزند. [۶۱،۶۲]. در جدول (۴-۴) مشتریان فعال فروشگاه نار بلا که الگوی خرید بدست آمده از آنها نشان داده است.

جدول (۴-۴) محصولات مورد علاقه هر مشتری

tv-led	tv-led	tv-plasma	tv-crt	tv-home	sys-audio	cd-dvd	rec-digital	yakh-tak	frizer-tak	sidebyside	yakh-friz-b	yakh-friz-p	yakh-friz-dogholoo
d	d	d	d	d	d	d	d	d	d	d	d	d	d
		1	1		1				1		1		
1	1	1		1	1						1		
1		1				1		1					
	1	1		1				1					1
1	1	1		1	1								1
1		1				1		1			1		
	1	1		1				1					
1	1	1		1	1								

و قسمتی از قوانین وابستگی انجام شده بر روی محصولات مورد علاقه مشتریان فعال در شکل (۶-۴) نشان داده شده است. دو فاکتور supp و conf به ترتیب مقدار پشتیبان و اطمینان قانون محصولات می باشد.

Rules	Supp	Conf
▲ tv-lcd		
tv-lcd -> tv-plasma	0.625	1.000
▲ frizer-tak		
frizer-tak -> tv-plasma	0.625	1.000
▲ tv-plasma		
tv-plasma -> tv-lcd	0.625	0.625
tv-plasma -> tv-led	0.750	0.750
tv-plasma -> tv-led tv-home	0.750	0.750
tv-plasma -> tv-home	0.750	0.750
tv-plasma -> frizer-tak	0.625	0.625
▷ tv-home		
▷ tv-led		
▲ tv-home		
tv-home -> tv-led tv-plasma	0.750	1.000
tv-home -> tv-led	0.750	1.000
tv-home -> tv-plasma	0.750	1.000
▷ tv-plasma		
▷ tv-led		
▲ tv-led		
tv-led -> tv-plasma	0.750	1.000
tv-led -> tv-plasma tv-home	0.750	1.000
tv-led -> tv-home	0.750	1.000
▷ tv-plasma		
▷ tv-home		

شکل (۶-۴) قوانین بدست آمده از محصولات

در اینجا قوانین ایجاد شده را به طور دقیق در دسته ای از خوشه های مناسب (که قبلاً ایجاد کردیم) نسبت می دهیم. این کار با میزان بدست آمده از فاکتور ارزیابی خوشه برای ارائه محصول انجام می گیرد. یعنی هر چه مقدار فاکتور ارزیابی خوشه بیشتر باشد می توان دو مقدار فاکتور قانون را بیشتر در نظر گرفت. در این مطالعه قوانین با پشتیبان ۶۰ و با اطمینان ۱۰، ۱۰۰ قانون برای خوشه اول محصولات ارائه می شوند. برای خوشه دوم با پشتیبان ۷۰ و با اطمینان

۱۰۰، ۸ قانون ارائه می شود. و برای خوشه هایی که نويز را اختيار کردند. هيچگونه قانونی تعلق نمی گیرد.

۸-۲ نتیجه کلی از کار پیشنهادی

در این تحقیق با ارائه الگوریتم خوشه بندی چالش هایی از جمله تشخیص تعداد خوشه بندی، گیر کردن بهینه محلی به منظور دقت در عملکرد را برطرف نموده ایم. در این مساله با مطالعه رویکردهای متفاوت و نادیده گرفتن بر روی نرخ روش های مورد استفاده بر روی الگوریتم ها ژنتیک خوشه بندی، بر آن شده ایم، که با حل کردن نرخ ها، بتوانیم در جهت حل مساله حرکت کنیم. و در ادامه نیز از دید جنبه کاربردی خوشه بندی به طور مناسب فرمول دهی شده است. تا مساله دقت چه در زیر ساختی و چه در جنبه کاربردی حل شود. این کار بروی مشتریان فروشگاه لوازم خانگی ناریلا انجام گرفته شده است. و پس از آن کشف قوانین بر روی محصولات مشتری اعمال شده، تا با دادن شرایط مناسب بروی قوانین، ارائه دقیقی از محصولات نسبت به مشتریان انجام گیرد. درواقع با حل عوامل پارامترهای درگیر مساله، مساله را حل نموده ایم. با حل کردن مسائل دقت در زیر ساختی و کاربردی، درجهت حرکت اهداف مدیریت سازمان ها بوده است. نیل به اهداف سازمانی بستگی تام به تعریف و تعیین نیازها و خواسته های بازارهای هدف و تأمین رضایت مشتری به نحوی مطلوب تر و مؤثرتر از رقیب دارد. در دنیای رقابتهای فزاینده امروز، شرکتهایی موفق خواهند بود، که رضایت مشتریان خود را بیشتر تأمین کنند. شرکتهایی که صرفاً به دنبال فروش کوتاه مدت نبوده و کسب رضایت مشتریان به صورت بلند مدت می باشند. که این کار از طریق ارائه کالاها و خدمات مورد علاقه مشتریان قرار می دهند.

۵- نتیجه گیری و آینده نگری

۵-۱ نتیجه گیری کلی

فصل پنجم

نتیجه گیری و کارهای آتی

خوشه بندی کلی: این گزارش به دیگر گزارش ها پیوسته قرار گرفته است که عملکرد خوب نسبت به آنها داشته است و پس از آن چند کاربردی را مورد بحث قرار دادیم. که از آن پس از فرمول بندی رفتار مشتریان و داده های مدل از محصولات، توانستیم به مشتریان خوشه بندی شده، محصولات مورد علاقه شان را پیشنهاد دهیم و نتیجه مشاهده کردیم که فرمول پیشنهادی از خوشه بندی زیر ساختی تأثیر بسزایی داشته است.

۶- آرایش بین دریاچه روندهای آینده نرم افزارهای مدیریت روابط مشتری

خوشه بندی کلی: که به موضوعات مختلف گشوده شده است. از جمله خوشه بندی است. بدون طرح XML، خوشه بندی بر روی تشخیص نحوه راه رفتن می باشد.

۵- نتیجه گیری و آینده نگری

۵-۱ نتیجه گیری کلی

در تحقیق حاضر، با مطالعه مقالات حال، توانستیم با دیدن نقاط قوت و ضعف در کارهایشان، به حوزه کاری تحقیقمان بکشانیم و با یک استراتژی جدید، دو رویکرد زیرساختی و کاربردی پیشنهاد داده شده است. الگوریتم خوشه بندی نمونه ای از کار زیرساختی می باشد که با حل کردن چالش های موجود در آن توانستیم مشتریان را با دقت خوشه بندی کنیم. این الگوریتم با دیگر الگوریتم ها مورد مقایسه قرار گرفته است که عملکرد خوبی نسبت به آنها داشته است. و پس از آن جنبه کاربردی را مورد بحث قرار دادیم، که آن نیز، از فرمول بندی رفتار مشتریان و با ارائه مدل از محصولات، توانستیم به مشتریان خوشه بندی شده، محصولات مورد علاقه شان را پیشنهاد دهیم. در نتیجه مشاهده کردیم که فرمول پیشنهادی در خوشه بندی زیر ساختی تاثیر بسزایی داشته است.

۵ پیشنهادی درباره روندهای آینده نرم افزارهای مدیریت روابط مشتری

حوزه خوشه بندی که به موضوعات مختلف کشانده شده است: از جمله خوشه بندی اسناد بدون طرح xml، خوشه بندی بر روی تشخیص نحوه راه رفتن و... می باشد.

از آنجا که خوشه‌بندی یک روش یادگیری بدون نظارت محسوب می‌گردد، در موارد بسیاری می‌تواند کاربرد داشته باشد. در بازاریابی (Marketing): دسته‌بندی مشتری‌ها به دسته‌هایی بر حسب رفتارها و نیازهای آنها از طریق مجموعه زیادی از ویژگی‌ها و آخرین خریدهای آنها. زیست‌شناسی (Biology): دسته‌بندی حیوانات و گیاهان از روی ویژگی‌های آنها. کتابداری: دسته‌بندی کتابها. نقشه‌رزی شهری (City-Planning): دسته‌بندی خانه‌ها بر اساس نوع و موقعیت جغرافیایی آنها. مطالعات زلزله‌نگاری (Earthquake studies): تشخیص مناطق حادثه‌خیز بر اساس مشاهدات قبلی. وب (WWW): دسته‌بندی اسناد و یا دسته‌بندی مشتریان به سایتها و ... در تشخیص گفتار (Speech Recognition): در ساخت کتاب کد از بردارهای ویژگی، در تقسیم کردن گفتار بر حسب گویندگان آن و یا فشرده‌سازی گفتار در تقسیم‌بندی تصاویر (Image Segmentation): تقسیم‌بندی تصاویر پزشکی و یا ماهواره‌ایدر تحقیق حاضر، سمت و سوی آن یک بخش‌بندی دقیقی از مشتریان بوده است که شالوده آن نرم افزار مدیریت روابط مشتری می‌باشد نرم افزارهای اجتماعی CRM در سال آینده به جریان اصلی بازار تبدیل خواهند شد. با افزودن پروفایل‌ها و اطلاعات اجتماعی مشتریان از روی شبکه‌های اجتماعی از قبیل توئیتر، LinkedIn و غیره به یکی از قابلیت‌های ضروری برای سازمان‌های فروشنده این نرم افزارها در سال ۲۰۱۲ تبدیل خواهد شد. توانمندی این سازمان‌ها برای تلفیق تعاملات اجتماعی مشتریان با نرم افزارهای مدیریت ارتباط با مشتری برای ارائه خدمات نوآورانه و موثر ضروری است، فروشنده‌گانی که این ملزومات را برآورده نکنند، امتیاز کافی را در این فضای رقابتی به دست نخواهند آورد. نرم افزارهای اجتماعی CRM جریان اصلی بازار خواهد بود و افزودن پروفایل‌ها و اطلاعات اجتماعی مشتریان از روی شبکه‌های اجتماعی به یکی از قابلیت‌های ضروری سازمان‌های فروشنده این نرم‌افزارها تبدیل خواهد شد. استفاده از روش‌های مبتنی بر سرگرمی (Gamification) یا به عبارت دیگر استفاده از سازوکار بازی و سرگرمی در فعالیتهای بازاریابی برای جلب مشارکت مشتریان است که تکنیک‌های شناخت رفتار مشتری را به همراه دارد. و در استراتژی دیگر خوشه بندی،

پروفایل خریداران و مشتریان علاوه بر اینکه اطلاعاتی درباره شغل و علایق را بدست می آورد ، از این به بعد نشان خواهد داد که شما به چه کسی اعتماد دارید.

بر این باور هستیم که درک و شناخت هویت اجتماعی خریداران و مشتریان در سال پیش رو اهمیت بیشتری به خود خواهد گرفت. سیستم‌های CRM نیز رفته رفته در حال اضافه کردن این اطلاعات به پروفایل مشتریان خود هستند و این خود فرصتی است برای درک اینکه اعتماد مردم و استفاده از آن به عنوان یک اهرم بازاریابی رو به افزایش است. روش‌های دیگر راهکارها داده‌های پویای برگرفته از منابع ساختاریافته و غیرساختاریافته را با برخی از اطلاعات هوشمند و پیشگویانه ترکیب خواهد کرد و نوعی از دانش و آگاهی را تولید خواهد کرد که بینش مشتریان را افزایش دهد.

«منابع»

- [۱] فیلیپ کاتلر دایرالمعارف بازاریابی از A تا Z ترجمه: عبدالحمید ابراهیمی ، و دیگران تهران : انتشارات آن.
- [۲] علیرضا مهدی ۱۳۸۶، "مقدمه‌ای بر الگوریتم ژنتیک و کاربردهای آن"، تهران: انتشارات ناقوس.
- [۳] مسعودیان سولمازو استکی افسانه ۱۳۸۶، "طراحی جدول زمانبندی خودکار با استفاده از الگوریتم های ژنتیک"، پایان نامه کارشناسی دانشگاه اصفهان.
- [4] C.Rygielski, A.J Wang, D.C. Yen., "Data mining techniques for customer relationship management", *Technology in Society* 24 ,2002 ,483–502.
- [5] M.A Farajian, S.Mohammadi. "Mining the Banking Customer Behavior Using Clustering and Association Rules Methods", *Internattiionall Journall off Industriiall Engineering & Producttiion Research*, December 2010, Volume 21, Number 4, pp. 239-245.
- [6] U.Fayyad;M. Gregory Piatetsky-Shapiro, and P.Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 2008,12-17.
- [7] S. Chakrabarti, M. Ester, U. Fayyad, J. Gehrke, Jiawei Han., "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2011-10-28.
- [8] N.C Romano , J.Fjermestad , "Electronic Commerce Customer Relationship Management: A Research Agend", Volume 4, Numbers 2-3,2003, 233-258
- [9] P. Berkhin, "A Survay of Clustering Data mining Techniques", 2002.
- [10] H.G Bennett, M.D Kuhn, " The Emergence of Electronic Customer Relationship Management", 2002. The Telecommunication Review, Vol. 13, p.91–96.
- [11] Cho, Y., Im, I., Hiltz, R., and Fjermestad, J. "Causes and outcomes of online customer complaining behavior: implications for customer relationship management (CRM)", *Proceedings of the 7th Americas Conference on Information Systems*, 2001, pp.900-907.

[12] P.S Bradley,U. M Fayyad, "Refining Initial Points for K-Means Clustering", In J. Shavlik, editor, Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann, San Francisco.1998,pp.91-99

[13] N. Balac; "Introduction to data mining", 2006.

[14] B. Fernandez/et. AL, "Knowledge management"; cho.12, 2004.

[15] M.Kantardzic."Data Mining: Concepts, Models, Methods, and Algorithms". John Wiley & Sons. ISBN 0471228524. OCLC 50055336,2003,pages 343

[16] A.Guazzelli, W.C Lin, T.Jena." PMML in Action: Unleashing the Power of Open Standards for Data Mining and Predictive Analytics". CreateSpace, 1452858268,2010,pages 188.

[17] The Data Mining Group (DMG). The DMG is an independent, vendor led group which develops data mining standards, such as the Predictive Model Markup Language (PMML).

[18] PMML(Predictive Model Markup Language) Project Page

[19] A. Guazzelli, M. Zeller, W.C Lin, G. Williams. "PMML: An Open Standard for Sharing Models". The R Journal, vol 1/1, 2009.

[20] J. Li," Data Mining - Clustering by Mixture Models",Department of Statistics The Pensilvania Stat University.

[21] D. Michie, D.J. Spiegelhalter, C.C. Taylor, "Machine Learning, Neural and Statistical Classification", 1994.

[22] J.Holland,"Adaptive In Natural and Artificial Systems".University of Michigan Press,1975.

[23] M.Vose, G.Liepins,"Punctuated Equilibria in Genetic Search.Complex System",1991,5:31-44.

[24] D.Goldberg, "A Note on Boltzmann Tournament Selection for Genetic Algorithms and Population-oriented Simulated Annealing".TCGA 90003, Engineering Mechanics , Univ.Alabama,1990.

[25] H. B Amor, A. Rettinger, "Intelligent exploration for genetic algorithms: Using self-organizing maps in evolutionary computation", In proceedings of the conference on Genetic and evolutionary computation, 2005, pp. 1531–1538.

- [26] D.E Goldberg ".Genetic Algorithms in Search, Optimization and Machine Learning", Kluwer Academic Publishers, Boston, MA. 1989.
- [27] D.E Goldberg,"The Design of Innovation: Lessons from and for Competent Genetic Algorithms", Addison-Wesley, Reading, MA.2002.
- [28] D.B Fogel, "Evolutionary Computation: Toward a New Philosophy of Machine Intelligence", IEEE Press, Piscataway, NJ. Third Edition,2006.
- [29] L.H Randy, E.H Ue, "Paractical Genetic Algorithm" (2nd ed) ,USA:Wiley 2004 .
- [30] F. Marki, M.Vogel, M,Fischer, " Process Plan Optimization Using a Genetic Algorithm ", 2006, pp. 528-531.
- [31] R. Perzina,"Solving the University Timetabling Problem with Optimized Enrolment of Students by a Parallel Self-adaptive Genetic Algorithm", PATAT , ISBN 80-210-3726-1,2006, pp. 264–280.
- [32] Y.G. Liu, K.F. Chen, X.M. Li, " A hybrid genetic based clustering algorithm", in: Proceedings of the 3rd International Conference on Machine Learning and Cybernetics, Shanghai, IEEE, 2004, pp. 1677–1682.
- [33] y. Xu,D.c Yen,B Lin, D.C Chou,"Adopting customer relation management technology", industrial Management & Data Systems 102/8 ,2002, 442-452.
- [34] A.K.Jain, R.C. "DubesAlgorithm for Clustering Data" .Prentice-Hall,Englewood Cliffs,NJ. 1998.
- [35] L.Kuofman and P.j.Russeeuw."Finding Groups in Data :An Introduction to Cluster Analysis".Wiley,New York. . 1990.
- [36]F.Murtagh."MultidimentionalClusteringAlgorithms".Physicya-Verlag,Vienna,Austria. 1985
- [37]C.Olson,"Parallel algorithms for hierarchical clustering.", Appeared in Parallel Parallel Computing, 1995, 21:1313-1325

- [38] A. R. Web, "Statistical Pattern Recognition", John Wiley & Sons, 2002.
- [39] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1967, 1:281-297.
- [40] Q. He, "A Review of Clustering Algorithms as Applied in IR", Graduate School of Library and Information Science University of Illinois at Urbana-Champaign, 1999.
- [41] A. Baraldi, P. Blonda, "A Survey of Fuzzy Clustering Algorithms for Pattern Recognition—Part I and II", IEEE Transaction systems, Man, And Cybernetics-part b: cybernetics, vol. 29, No. 6, 1999.
- [42] N. Memarsadeghi, D. M. Mount, N. S. Netanyahu, J. LeMoigne, "A Fast Implementation of the ISODATA Clustering Algorithm", International Journal of Computational Geometry and Applications. 2007, pp. 71-103.
- [43] N. Memarsadeghi, D. M. Mount, N. S. Netanyahu, J. LeMoigne, "A Fast Implementation of the ISODATA Algorithm", International Geoscience and Remote Sensing Symposium (IGARSS'03), Toulouse, France, July 21-25, 2003.
- [44] L. Fuyan, C. Chouyong, and L. Shaoyi, "An Improved Genetic Approach", Proceeding of ICSC IEEE, vol. 2, 2005, pp. 641-644.
- [45] S. Yang, Y. Li, "k-means Optimization Study on k Value of K2means Algorithm", System Engineering Theory and Application, 2006, pp. 97-101.
- [46] M. Laszlo, S. Mukherjee, "A genetic algorithm using hyper-quadtrees for low-dimensional K-means clustering", IEEE Trans. Pattern Anal. Mach. Intell. 2006, pp. 33-543.
- [47] Y. Liu, X. Wu, Y. Shen, "Automatic clustering using genetic algorithms", Applied Mathematics and Computation 218, 2011, pp. 1267-1279.
- [48] K. G. Srinivasa, K. R. Venugopal, L. M. Patnaik, "A self-adaptive migration model genetic algorithm for data mining applications", Information Science, 2007, 117, pp. 4295-4313.
- [49] H. He, Y. Tan, "A two-stage genetic algorithm for automatic clustering", Neurocomputing 81, 2012, pp. 49-59.
- [50] J. Donio, P. Massari, G. Passiante, "Customer satisfaction and loyalty in a digital environment: an empirical test", Journal of Consumer Marketing, Vol. 23 Iss: 7, 2006, pp. 445 - 457.

[51] Bing fang , Shoufeng,Data mining technology and its application in CRM of commercial banks.First international workshop on database technology and application 2009 IEEE, DOI 10,1109,DBTA.2009,128.

[52] M. Khajvand and M. Jafar Tarokh." Analyzing Customer Segmentation Based on Customer Value Components (Case Study: A Private Bank)"
(Technical note). Journal of Industrial Engineering, University of Tehran, Special Issue, 2011, PP. 79-93

[53] D.R Liu,Ya-Yueh Shih., "Integrating AHP and datamining for product recommendation based on customer lifetime value",information & Mangement 42,2005,pp: 387-400.

[54] H.J Chang, L.P Hung , C.L Ho, " An anticipation model of potential customers' purchasing behavior based on clustering analysis and association rules analysis", Expert Systems with Applications 32,2007,pp: 753-764

[55] j Curry , A Curry."the customer marketing method :How to implement and profit from Customer Relationship Management", ISBN-10: 0684839431, Free Press March 27, 2000,pages 256.

[56] V. Aggelis,W. Piraeus,"customer clustering using RFM analysis".Citcseer ,2005 ,pp:1-5

[57] C. West, S. MacDonald, P. Lingras1, G Adams., " Relationship between Product Based Loyalty and Clustering based on Supermarket Visit and Spending Patterns", International Journal of Computer Science & Applications © 2005 Technomathematics Research Foundation , Vol. II, No. II, ,2005, pp. 85 – 100.

[58] R.Sivaraj, Dr.Travichandran, "A Review of Selection Methods in Genetic Algorithm", International Journal of Engineering Science and Technology (IJEST). ISSN:0975-5462 , Vol. 3, Issue.5,2011,pp:3792-3797.

[59] F. Kovács, C. Legány, A. Babos, "Cluster Validity Measurement Techniques", Department of Automation and Applied Informatics, Budapest University of Technology and Economics, 2003.

[60] B.Ozdden, S.Ramaswamy, and A.Sillberschatz, "Cyclic association rules," In proceedings of the 14th International Conference on Data Engineering, Feb 1998 ,pp: 412-421,

[61] S. Ramaswamy, S.Mahajan, and A.Sillberschatz, "On the discovery of interesting patterns in association rules," In Proceedings of the 1998 International Conference on Very Large Data Base, August 1998,pp:368-379.

[62] C. Bettini, X. Sean Wang R: " Time Granularies in databases , Data Mining , and Temporal reasoning", Springer-Verlag ,ISBN 3-540- 66997-3, July 2000. pages 230 .

[63] <http://archive.ics.uci.edu/ml/datasets/Iris>

مجموعه داده iris

Object	Coordinate of object			
Number	x1	x2	x3	x4
1	5.1	3.5	1.4	0.2
2	4.9	3	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.4	0.2
5	5	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.8	3.4	1.4	0.3
8	5	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	1.5	0.1
11	5.4	3.7	1.5	0.2
12	5.0	3.4	1.4	0.3
13	4.7	3	1.4	0.3
14	4.3	3	1.1	0.1
15	5.0	4	1.2	0.2
16	5.2	4.4	1.5	0.4
17	5.4	3.6	1.3	0.4
18	5.1	3.5	1.4	0.3
19	5.7	3.8	1.7	0.3
20	5.1	3.8	1.5	0.3
21	5.4	3.4	1.7	0.2
22	5.1	3.7	1.6	0.4
23	4.9	3.5	1	0.2
24	5.1	3.3	1.7	0.5
25	4.8	3.4	1.0	0.2
26	5	3	1.6	0.2
27	5	3.4	1.6	0.4
28	5.2	3.5	1.5	0.2

پیوست یک

مجموعه داده iris

object	Coordinate of object			
Number	x1	x2	x3	x4
1	5.1	3.5	1.4	0.2
2	4.9	3	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	1.5	0.1
11	5.4	3.7	1.5	0.2
12	4.8	3.4	1.6	0.2
13	4.8	3	1.4	0.1
14	4.3	3	1.1	0.1
15	5.8	4	1.2	0.2
16	5.7	4.4	1.5	0.4
17	5.4	3.9	1.3	0.4
18	5.1	3.5	1.4	0.3
19	5.7	3.8	1.7	0.3
20	5.1	3.8	1.5	0.3
21	5.4	3.4	1.7	0.2
22	5.1	3.7	1.5	0.4
23	4.6	3.6	1	0.2
24	5.1	3.3	1.7	0.5
25	4.8	3.4	1.9	0.2
26	5	3	1.6	0.2
27	5	3.4	1.6	0.4
28	5.2	3.5	1.5	0.2

29	5.2	3.4	1.4	0.2
30	4.7	3.2	1.6	0.2
31	4.8	3.1	1.6	0.2
32	5.4	3.4	1.5	0.4
33	5.2	4.1	1.5	0.1
34	5.5	4.2	1.4	0.2
35	4.9	3.1	1.5	0.1
36	5	3.2	1.2	0.2
37	5.5	3.5	1.3	0.2
38	4.9	3.1	1.5	0.1
39	4.4	3	1.3	0.2
40	5.1	3.4	1.5	0.2
41	5	3.5	1.3	0.3
42	4.5	2.3	1.3	0.3
43	4.4	3.2	1.3	0.2
44	5	3.5	1.6	0.6
45	5.1	3.8	1.9	0.4
46	4.8	3	1.4	0.3
47	5.1	3.8	1.6	0.2
48	4.6	3.2	1.4	0.2
49	5.3	3.7	1.5	0.2
50	5	3.3	1.4	0.2
51	7	3.2	4.7	1.4
52	6.4	3.2	4.5	1.5
53	6.9	3.1	4.9	1.5
54	5.5	2.3	4	1.3
55	6.5	2.8	4.6	1.5
56	5.7	2.8	4.5	1.3
57	6.3	3.3	4.7	1.6
58	4.9	2.4	3.3	1
59	6.6	2.9	4.6	1.3
60	5.2	2.7	3.9	1.4
61	5	2	3.5	1
62	5.9	3	4.2	1.5
63	6	2.2	4	1
64	6.1	2.9	4.7	1.4
65	5.6	2.9	3.6	1.3
66	6.7	3.1	4.4	1.4

67	5.6	3	4.5	1.5
68	5.8	2.7	4.1	1
69	6.2	2.2	4.5	1.5
70	5.6	2.5	3.9	1.1
71	5.9	3.2	4.8	1.8
72	6.1	2.8	4	1.3
73	6.3	2.5	4.9	1.5
74	6.1	2.8	4.7	1.2
75	6.4	2.9	4.3	1.3
76	6.6	3	4.4	1.4
77	6.8	2.8	4.6	1.4
78	6.7	3	5	1.7
79	6	2.9	4.5	1.5
80	5.7	2.6	3.5	1
81	5.5	2.4	3.8	1.1
82	5.5	2.4	3.7	1
83	5.8	2.7	3.9	1.2
84	6	2.7	5.1	1.6
85	5.4	3	4.5	1.5
86	6	3.4	4.5	1.6
87	6.7	3.1	4.7	1.5
88	6.3	2.3	4.4	1.3
89	5.6	3	4.1	1.3
90	5.5	2.5	4	1.3
91	5.5	2.6	4.4	1.2
92	6.1	3	4.6	1.4
93	5.8	2.6	4	1.2
94	5	2.3	3.3	1
95	5.6	2.7	4.2	1.3
96	5.7	3	4.2	1.2
97	5.7	2.9	4.2	1.3
98	6.2	2.9	4.3	1.3
99	5.1	2.5	3	1.1
100	5.7	2.8	4.1	1.3
101	6.3	3.3	6	2.5
102	5.8	2.7	5.1	1.9
103	7.1	3	5.9	2.1
104	6.3	2.9	5.6	1.8

105	6.5	3	5.8	2.2
106	7.6	3	6.6	2.1
107	4.9	2.5	4.5	1.7
108	7.3	2.9	6.3	1.8
109	6.7	2.5	5.8	1.8
110	7.2	3.6	6.1	2.5
111	6.5	3.2	5.1	2
112	6.4	2.7	5.3	1.9
113	6.8	3	5.5	2.1
114	5.7	2.5	5	2
115	5.8	2.8	5.1	2.4
116	6.4	3.2	5.3	2.3
117	6.5	3	5.5	1.8
118	7.7	3.8	6.7	2.2
119	7.7	2.6	6.9	2.3
120	6	2.2	5	1.5
121	6.9	3.2	5.7	2.3
122	5.6	2.8	4.9	2
123	7.7	2.8	6.7	2
124	6.3	2.7	4.9	1.8
125	6.7	3.3	5.7	2.1
126	7.2	3.2	6	1.8
127	6.2	2.8	4.8	1.8
128	6.1	3	4.9	1.8
129	6.4	2.8	5.6	2.1
130	7.2	3	5.8	1.6
131	7.4	2.8	6.1	1.9
132	7.9	3.8	6.4	2
133	6.4	2.8	5.6	2.2
134	6.3	2.8	5.1	1.5
135	6.1	2.6	5.6	1.4
136	7.7	3	6.1	2.3
137	6.3	3.4	5.6	2.4
138	6.4	3.1	5.5	1.8
139	6	3	4.8	1.8
140	6.9	3.1	5.4	2.1
141	6.7	3.1	5.6	2.4
142	6.9	3.1	5.1	2.3

143	5.8	2.7	5.1	1.9
144	6.8	3.2	5.9	2.3
145	6.7	3.3	5.7	2.5
146	6.7	3	5.2	2.3
147	6.3	2.5	5	1.9
148	6.5	3	5.2	2
149	6.2	3.4	5.4	2.3
150	5.9	3	5.1	1.8

۱- تغییر location مربوط به مجموعه داده‌های باشد تغییر cluster مربوط به نگهداری کمترین فاصله کلیه‌های می باشد تغییر cluster بر مجموعه داده به همراه شماره حائلی از اطلاعات آنها می باشد.

global LOCATION
global index
global cluster

۲- دستور ورود داده‌های global به نام location می باشد
LOCATION=insertval('loc','location')
n = 3
X=LOCATION

۳- برای این هر یک از اشیاء واریاس کلی حرکت می شود. در نهایت الحروف بعدی نسبت می آید.
for i = 1:length(X)
X(i,1)=abs(X(i,2)-LOCATION(i,2))-LOCATION(i,4)
LOCATION(i,3)=
end
w=length(X)/2
w=length(X)/2

۴- محاسبه‌های واریاس که تعداد حالت‌ها را مشخص می کند.
if (w<0.01)
T=20
else
if (w<0.01)
T = 40
else
if (w<0.01)
T=60
else
T=80

پروژه ۲ فرمات

پیوست دو

کد الگوریتم پیشنهادی

۱- متغیر location مربوط به مجموعه داده می باشد و متغیر index مربوط به نگهداشتن کمترین فاصله اقلیدوسی می باشد. متغیر cluster نیز مجموعه داده به همراه شماره حالتی از انتخاب ژنها می باشد.

```
global LOCATION
global index
global cluster
```

۲- دستور ورود داده های excel به متغیر location می باشد

```
LOCATION=(xlsread('Clust','Location'));
x = 0
X=LOCATION
```

۳- واریانس هر یک از اشیاء و واریانس کلی گرفته می شود. در نهایت انحراف معیار بدست می آید.

```
For i=1:size(LOCATION,1)
varrow(i,1)=var([LOCATION(i,2) LOCATION(i,3) LOCATION(i,4)
LOCATION(i,4)] )
end
varkol=var(varrow,1)
v=sqrt(varkol)
```

۴- معیار های واریانس که تعداد حالت ها را مشخص می کند.

```
if (v<0.01)
    T=20
else
    if (v<0.5)
        T = 40;
    else
        if (v<1)
            T=60;
        else
```

```

(v<1.5)
    T=80;
end
end
end
end

```

۵- تشخیص اندازه و حالت های خوشه.

```

cluster(1:150,1)=20
num=T*num;
PopSize=input('Population size:');

```

۶- فراخوانی تابع ارزیابی و تنظیمات تقاطع ۰.۹ و جهش نرخ ۱ و در نهایت اجرای عملیات ژنتیکی انجام می گیرد.

```

FitnessFcn = @Cluster;
numberOfVariables = num;
LOCATION=(xlsread('Clust3','Location'));
Options=gaoptimset('PopInitRange',[0;1],'PopulationSize',PopSize,'SelectionFcn',@selectionroulette,'CrossoverFcn',{@crossoverscattered},'MutationFcn',{@mutationadaptfeasible,0.9});
[x,fval] = ga(FitnessFcn,numberOfVariables,Options);

```

۷- محاسبه فاصله اقلیدوسی هر یک از اشیای مجموعه داده با حالت ها مشخص شده.

```

for i=1:size(LOCATION,1)
for j=1:(size(x,2)/4)
distances(i,j)=sqrt((LOCATION(i,2)-x(j))^2+(LOCATION(i,3)-x(size(x,2)/4+j))^2+(LOCATION(i,4)-x(2*size(x,2)/4+j))^2+(LOCATION(i,5)-x(3*size(x,2)/4+j))^2);
end
end

```

۸- مینم فاصله ها در داخل index قرار می گیرد (به منظور مشاهده).

```

for i=1:size(LOCATION,1)
index(i)=min(distances(I,:));
end

```

۹- مقدار دهی و مرتب کردن شماره حالت در درون statenumsort.

```

clustersort(:,1)=cluster(:,2)
statenumsort(:,1)=sort(clustersort,1);
state=0
f=1

```

۱۰- پیدا کردن شماره حالت برای pointhalat.

```

for i=1:size(LOCATION,1)
if state~=statenumsort(i,1)
pointhalat(1,f)=statenumsort(i,1)

```

```

state=pointthalat(1,f)
f=f+1;
end
end

```

۱۱- تعداد خوشه مشخص می شود.

```
Nc=f-1
```

۱۲- اشیا با شماره حالت یکسان در یک خوشه دسته بندی می شوند.

```

for i=1:size(LOCATION,1)
    X(i,6)=cluster(i,2)
end

s=1
for l=1:Nc
    for i=1:size(LOCATION,1)
        if X(i,6)== pointthalat(1,1)
            class{l}(s,:)=X(i,:);
            s=s+1
        end
    end

    en
    if s==2
        r(1,:)=class{l}(s-1,:)
        class{l}(s,:)=r(1,:)
    end
    s=1
end
maxcluster=0;
max=0
j=0

```

۱۳- محاسبه شاخص دون برای دو خوشه.

```

if (Nc==2)
    clust1=class{1}(:, :)
    clust2=class{2}(:, :)
    point(1,1)=numel(clust1)/6
    point(1,2)=numel(clust2)/6
    clust2
    maxcluster1=0
    maxcluster2=0
    mincluster12=10

```

۱۴- محاسبه ماکزیمم فاصله نقاط درون خوشه اول.

```

for i=1:point(1,1)-1
    for j=i+1:point(1,1)

```



```

        max2=sqrt((clust1(i,2)-clust1(j,2))^2+(clust1(i,3)-
        clust1(j,3))^2+(clust1(i,4)-clust1(j,4))^2+(clust1(i,5)-
        clust1(j,5))^2)
        if max2>maxcluster1
            maxcluster1=max2
        end
    end
end

```

۱۵- محاسبه ماکزیمم فاصله نقاط درون خوشه دوم.

```

for i=1:point(1,2)-1
    for j=i+1:point(1,2)
        max2=sqrt((clust2(i,2)-clust2(j,2))^2+(clust2(i,3)-
        clust2(j,3))^2+(clust2(i,4)-clust2(j,4))^2+(clust2(i,5)-
        clust2(j,5))^2)
        if max2>maxcluster2
            maxcluster2=max2
        end
    end
end

```

۱۶- محاسبه مینیمم فاصله اقلیدوسی نقاط خوشه اول و دوم.

```

for i=1:point(1,1)
    for j=1:point(1,2)
        min2=sqrt((clust1(i,2)-clust2(j,2))^2+(clust1(i,3)-
        clust2(j,3))^2+(clust1(i,4)-clust2(j,4))^2+(clust1(i,5)-
        clust2(j,5))^2)
        if min2<mincluster12
            mincluster12=min2
        end
    end
end

```

۱۷- گام ۱۶ بر گام ۱۵ و ۱۴ تقسیم میشود و مقدار مینیمم به عنوان نتیجه ارزیابی خوشه نشان داده

می شود.

```

dxyz(1,1)= mincluster12
dxy(1,1)=dxyz(1,1)/maxcluster1
dxy(1,2)=dxyz(1,1)/maxcluster2
validation=min(dxy(1,:));
validation
end

```

۱۸- برای سه خوشه، چهارو ... به همین شکل نوشته شده است.

۱۹- تابع ارزیابی که مقدار برگشتی Z را دارد.

```

function z=Clust(x)
global LOCATION
global cluster
global varrow
global medrow

```

```

global zaribtaghirsh
global zaribtaghirx
global distances
global l
global w
global tarsim
z=0;

```

۲۰- محاسبه فاصله اقلیدوسی اشیا با هر یک از حالت های X و محاسبه مینیمم آنها به عنوان خروجی

```

varx=var(x)
medx=mean(x)
l=l+1
zaribtaghirx=sqrt(varx)/medx

for i=1:size(LOCATION,1)
for j=1:(size(x,2)/4)
distances(i,j)=sqrt((LOCATION(i,2)-x(j))^2+(LOCATION(i,3)-
x(size(x,2)/4+j))^2+(LOCATION(i,4)-
x(2*size(x,2)/4+j))^2+(LOCATION(i,5)-
x(3*size(x,2)/4+j))^2+(zaribtaghirsh(i,1)/zaribtaghirx);
end
min_distance=min(distances(i,:));
z=z+min_distance;
end

```

۲۱- ترسیم نسل ها

```

tarsim(w)=z;
w=w+1

```

۲۱- نگهداری مقدار مینیمم و شماره حالت X

```

for i=1:size(LOCATION,1)
for j=1:(size(x,2)/4)
if cluster(i,1)>distances(i,j)
cluster(i,2)=j
cluster(i,1)=distances(i,j)
end
end
end

```

Abstract

Nowdays, valid data extraction process . Understandable from the unknown And reliability of large database and its use in decision making is used in commercial activities. In the present study provide an efficient approach to clustering customers using clustering based on genetic optimization is performed automatically . And then offer Concepts of customer categories Using data mining tools orange Customers' favorite products Is modeling.

Uil



Islamic Azad University
Sari Branch

A Master's Thesis To Get M.cS«Degree
Major : Computer Science

Title:

Loyal customer by clustering algorithm based on
optimal and behavior customer prospect

Thesis Advisor:

Homayoon Motameni (Ph.D)

Consulting Advisor:

Javad vahidi (Ph.D)

By:

aref yelghi

Summer 2012



Islamic Azad University

Sari Branch

A Master's Thesis To Get M.cS Degree

Major : Computer Science

Title :

**Loyal customer by clustering algorithm based on
optimal and behavior customer prospect**

Thesis Advisor :

Homayoon Motameni (Ph.D)

Consulting Advisor :

Javad Vahidi (Ph.D)

By :

Aref yelghi

Summer 2012