

طراحی مدل ترکیبی داده کاوی و تصمیم گیری چند معیاره (مورد مطالعه: بانک اطلاعات یارانه های مرکز آمار ایران)

عادل آذر^{۱*}، علیرضا مهدوی راد^۲، مرتضی موسی خانی^۳

۱- استاد، دانشگاه تربیت مدرس، گروه مدیریت صنعتی، تهران، ایران

۲- کارشناس ارشد، دانشگاه آزاد اسلامی، گروه مدیریت صنعتی، قزوین، ایران

۳- دانشیار، دانشگاه آزاد اسلامی، گروه مدیریت دولتی، قزوین، ایران

رسید مقاله: ۱ خرداد ۱۳۹۳

پذیرش مقاله: ۱۰ مهر ۱۳۹۳

چکیده

حجم بالای داده ها در سازمان ها و جامعه ی امروزی حاصل پیشرفت های چشم گیر بشر در علوم کامپیوتری است. دغدغه ی مدیران امروز دیگر کمبود داده ها نیست؛ بلکه استفاده ی درست از داده های موجود و تبدیل آن ها به اطلاعات و سپس به دانش برای اتخاذ درست ترین تصمیمات با درصد خطای پایین تر می باشد. تکنیک های داده کاوی و روش های تصمیم گیری چند معیاره در دهه های اخیر هر کدام به شکلی کمک رسان مدیران در عرصه ی تصمیم گیری بوده است؛ اما این روش ها همواره نقاط ضعفی داشته اند؛ که موجب جستجوی پیوسته محققان در طراحی و کاربرد روش های جدید تصمیم گیری شده است. در پژوهش حاضر سعی بر آن داریم که با ارایه ی مدلی ترکیبی از داده کاوی و روش های تصمیم گیری چند معیاره، نتایج هر دو روش را بهبود دهیم. در این راستا مدل های تصمیم گیری چند معیاره می تواند نقش تصمیم گیرندگان را در نتایج مدل های داده کاوی، پررنگ تر کرده، از طرفی استفاده از تکنیک های داده کاوی امکان اجرای روش های تصمیم گیری چند شاخصه را بر روی حجم بالایی از داده ها امکان پذیر سازد. پس از ارایه ی چارچوب پیشنهادی برای مدل ترکیبی تحقیق یک مطالعه موردی بر روی پایگاه داده ای مناسب صورت گرفته است تا میزان کاربردی بودن چارچوب پیشنهادی مشخص گردد. این مطالعه ی موردی مربوط به خوشه بندی خانوارهای یکی از استان های کشور و رتبه بندی خانوارها با استفاده از مدل ترکیبی پیشنهادی می باشد.

کلمات کلیدی: داده کاوی، تصمیم گیری چند شاخصه، روش ویکور، خوشه بندی.

۱ مقدمه

در دنیای امروز مساله ی تصمیم گیری از مسایل بسیار مهمی است که ذهن اکثر مدیران را در سطوح مختلف به خود مشغول کرده است. در حل مسایل تصمیم گیری روش های بسیاری وجود دارد که

* عهده دار مکاتبات

آدرس الکترونیکی: AZARA@modares.ac.ir

روش های تصمیم گیری چند معیاره و تکنیک های داده کاوی از روش های پر کاربرد در این عرصه می باشد. مسایل بهینه سازی و تصمیم گیری زندگی واقعی نیازمند ابزارهای بهینه سازی چند معیاره است. با توجه به این نیاز، روش های تصمیم گیری چند معیاره از محبوبیت گسترده ای در علوم مختلف برخوردار می باشد. با روند رو به رشد کاربرد مدل های تصمیم گیری چند معیاره، افزایش تعداد اجزای تشکیل دهنده، متغیرها، پارامترها، اعداد ثابت و اهداف موجود در این فرآیند، این روش ها را بسیار پیچیده می کند. نسل جدید ابزارهای تصمیم گیری این فرآیند را هر چه بیش تر خود کار کرده است؛ اما شروع فرآیند و تنظیم ارزش های اولیه ای ابزارهای شبیه سازی و همچنین شناسایی اهداف و متغیرهای ورودی مؤثر برای رسیدن به فضای طراحی کوچک تر همچنان پیچیده می باشد [۱]. برای نمونه می توان به تحقیق آقای کریمی شیرازی و همکاران در صنایع غذایی ایران اشاره کرد. در تحقیق مذکور از روش ترکیبی ANP، DEMATEL و VIKOR استفاده شده است که هر کدام از این روش ها در جای خود پیچیدگی های خاص خود را دارا می باشد [۲].

روش دیگری که در تصمیم گیری های امروزی بسیار پر کاربرد می باشد داده کاوی است. تکنیک های داده کاوی، فرآیند تشخیص الگوها و مدل های موجود در داده ها است، الگوهایی که معتبر، بدیع، بالقوه، مفید و کاملاً قابل فهم می باشد. به عبارت دیگر، هدف تکنیک های داده کاوی یافتن الگوهای جالب موجود در پایگاه داده ها است که در میان حجم عظیمی از داده ها مخفی است [۳]. هر چند کاربرد تکنیک های داده کاوی در دنیای نرم افزار به عنوان الگوهای هوشمند حل مسایل دسته بندی، خوشه بندی، و مدل های پیش بینی، همچنان روبه پیشرفت و افزایش دارد؛ اما باید گفت که استفاده از این تکنیک ها در حل مسایل تصمیم گیری مدیران آن چنان که انتظار می رفته کاربردی نشده است. شاید بتوان گفت این دور ماندن تکنیک های داده کاوی و مدیران از یکدیگر به دلیل دیدگاه خود کار سازی هر چه بیش تر الگوریتم های داده کاوی در شاخه ی هوش مصنوعی باشد. حال آنکه در تصمیم گیری های مدیریتی نیاز به اعمال نظر مدیران در فرآیند تصمیم گیری است. در این تحقیق به دنبال آن هستیم تا با کاربرد تکنیک های داده کاوی و ترکیب آن ها با روش های تصمیم گیری چند معیاره، مدلی ترکیبی ارائه کنیم که خروجی هریک از این روش ها در مسایل تصمیم گیری را بهبود داده و تا حد امکان مدلی کاربردی باشد. نوآوری این تحقیق در استفاده از داده کاوری برای پیش پردازش داده های ورودی مدل های تصمیم گیری چند معیاره و امکان کار با تعداد بالای داده ها در این مدل ها و از طرفی استفاده از مدل های تصمیم گیری چند معیاره در جهت اعمال نظر مدیران در مدل های داده کاوی می باشد. مدل ارائه شده در مرکز آمار ایران و برای خوشه بندی و رتبه بندی خانوارهای یکی از استان های کشور مورد آزمایش قرار گرفته است.

۲ مفاهیم

۲-۱ کشف دانش و داده کاوی

همان طور که الکترون ها و امواج موضوع اصلی مهندسی برق شد، داده ها، اطلاعات و دانش نیز موضوع اصلی حوزه ی جدیدی از تحقیق و کاربرد به نام کشف دانش و داده کاوی است. به طور کلی،

داده کاوی را می توان به صورت مجموعه ای از مکانیسم ها و تکنیک ها که از طریق نرم افزار، اطلاعات مخفی شده در داده ها را استخراج می کند تعریف کرد [۴]. رایج ترین و پرکاربردترین تعریف از کشف دانش و داده کاوی تعریفی است که به فایاد و همکاران نسبت داده شده است "فرآیند شناخت و شناسایی الگوهای معتبر، بدیع، مفید و در نهایت قابل فهم در داده ها" [۵].

پرکاربردترین استاندارد رسمی برای داده کاوی استاندارد کریسپ (CRISP) است. به طور کلی استاندارد کریسپ دارای ۶ مرحله است.



شکل ۱. فرآیند کلی داده کاوی بر اساس استاندارد کریسپ

۲-۱-۱ خوشه بندی و الگوریتم k میانگین

خوشه بندی به معنای "تقسیم داده ها" به گروه های مشابه است. داده ها بر اساس اصل حداکثر کردن شباهت داخل گروهی و حداقل کردن شباهت بین گروهی، خوشه بندی می شود [۶].

الگوریتم k میانگین پارامتر k را به عنوان ورودی گرفته، مجموعه n شی را به k خوشه افراز می کند. به طوری که سطح شباهت داخلی خوشه ها بالا بوده و سطح شباهت اشیای بیرون خوشه ها پایین باشد. شباهت هر خوشه نسبت به متوسط اشیای آن خوشه، سنجیده شده که این متوسط، مرکز خوشه نیز نامیده می شود. ورودی و خروجی این الگوریتم به صورت زیر است:

ورودی: k به عنوان تعداد خوشه ها و یک پایگاه داده شامل n شیء

خروجی: یک مجموعه از k خوشه که معیار مربع خطا را حداقل می کند.

در عمل، این الگوریتم یک روش هیوریستیک برای کاهش معیار مربع خطاست که در رابطه ی زیر آمده است:

$$E = \sum \sum |p - m_i|^2 \quad (1)$$

در این رابطه E مجموعه ی مربع خطا برای تمام اشیای پایگاه داده می باشد. p نقطه ای در فضا است که نمایانگر یک شیء می باشد و m_i میانگین خوشه ی C_i می باشد که نقطه ی p به آن متعلق است. (هم p و هم m_i چندبعدی هستند) [۶].

یکی از مهم ترین نقاط ضعف این روش این است که در برابر اغتشاشات و نقاط پرت حساس است؛ زیرا این داده ها به راحتی مراکز را تغییر می دهند و ممکن است نتایج مطلوبی حاصل نشود. به همین دلیل در این تحقیق سعی شده قبل از استفاده از این روش تا حد امکان نقاط پرت، حذف یا اصلاح شود. برای این منظور از الگوریتم آنومالی که ضمن خوشه بندی، نقاط پرت را هم مشخص می کند و روش های آماری استفاده شده است که در پیاده سازی مدل توضیح داده خواهد شد.

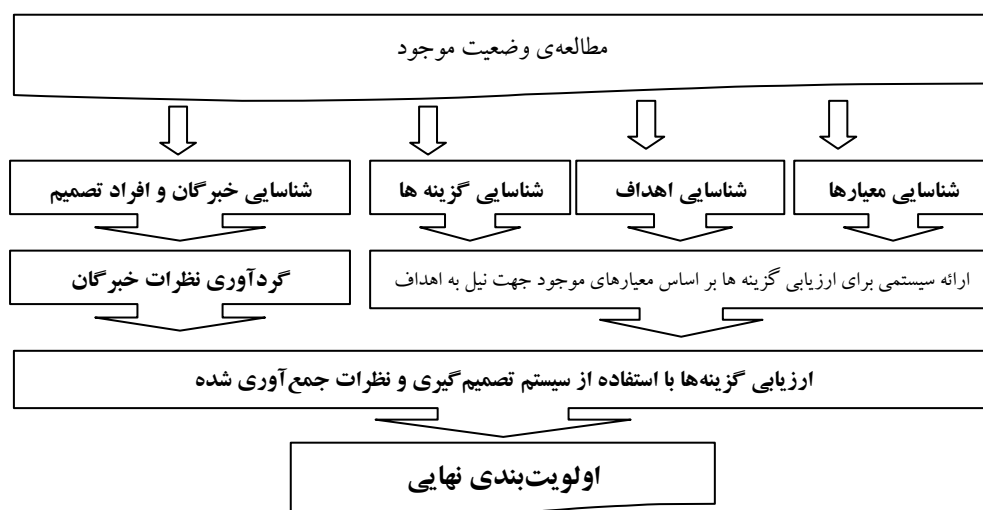
۲-۱-۲ دسته‌بندی

دسته‌بندی، فرآیند یافتن مدلی است که با تشخیص دسته‌ها یا مفاهیم داده می‌تواند دسته‌ی ناشناخته‌ی اشیای دیگر را پیش‌بینی کند. دسته‌بندی یک تابع یادگیری است که یک قلم داده را به یکی از دسته‌های از قبل تعریف‌شده نگاشت می‌کند. داده‌های موجود به دو قسمت آموزش و آزمون تقسیم می‌شوند. داده‌های آموزش برای یادگیری قواعد توسط سیستم استفاده می‌شود و داده‌های آزمون برای بررسی دقت دسته‌بندی و جلوگیری از بیش برداش به کار می‌روند. یکی از روش‌های متداول دسته‌بندی تکنیک درخت تصمیم و الگوریتم کارت می‌باشد [۴].

در واقع الگوریتم درخت تصمیم‌گیری با انتخاب آزمونی شروع می‌شود که بهترین جداسازی را برای دسته‌ها انجام دهد. در مراحل بعدی درخت هم، همین کار برای گره‌های بعدی با داده‌های کم‌تر صورت می‌گیرد تا بهترین قانون‌ها حاصل شود. درخت آن‌قدر بزرگ می‌شود تا دیگر نتوان جداسازی بهتری برای داده‌های گره انجام داد [۷].

۲-۲ تصمیم‌گیری چندمعیاره و روش ویکور

به‌طور کلی فرآیند تصمیم‌گیری چندمعیاره عبارت است از طریقه‌ی پیدا کردن وزن‌ها، رتبه‌ها و یا اهمیت یک مجموعه از فعالیت‌ها با توجه به تأثیر آن‌ها بر وضعیت موجود و هدف تصمیم. فرآیند کلی تصمیم‌گیری چندمعیاره در شکل (۲) نشان داده شده است [۸].



شکل ۲. فرآیند کلی تصمیم‌گیری چندمعیاره

۲-۲-۱ روش ویکور

ویکور یک روش تصمیم‌گیری چندمعیاره‌ی توافقی است که توسط آپریکوویچ و زنگو بر مبنای روش ال پی متریک توسعه یافته است [۹].

این روش می تواند یک مقدار بیشینه مطلوبیت گروهی برای اکثریت و یک کمینه تأثیر انفرادی برای مخالفت را فراهم نماید.

مراحل کاری روش ویکور

مراحل این روش شامل گامهای ذیل است [۹]:

۱. محاسبه مقادیر نرمال شده

۲. تعیین بهترین و بدترین مقدار

بهترین و بدترین هریک از مقادیر در هر معیار را شناسایی می کنیم و به ترتیب f_j^* و f_j^- می نامیم.

$$f_j^* = \text{Max } f_{ij}, i = 1, 2, \dots, m \quad (2)$$

$$f_j^- = \text{Min } f_{ij}, j = 1, 2, \dots, n \quad (3)$$

۳. تعیین وزن معیارها

اوزان معیارها باید برای بیان اهمیت روابط آنها محاسبه شده باشد. برای این منظور، در این تحقیق از روش تحلیل سلسله مراتبی استفاده شده است.

۴. محاسبه فاصله ی گزینه ها از راه حل ایده آل

این مرحله محاسبه فاصله ی هر گزینه از راه حل ایده آل و سپس حاصل جمع آنها برای ارزش نهایی بر اساس روابط ذیل است:

$$S_i = \sum_{j=1}^n w_j (f_j^* - f_{ij}) / (f_j^* - f_j^-) \quad (4)$$

$$R_i = \text{Max}_j [w_j (f_j^* - f_{ij}) / (f_j^* - f_j^-)] \quad (5)$$

جایی که S_i بیانگر نسبت فاصله ی گزینه i ام از راه حل ایده آل مثبت (بهترین ترکیب) و R_i بیانگر نسبت فاصله ی گزینه i ام از راه حل ایده آل منفی (بدترین ترکیب) می باشد.

۵. محاسبه مقدار ویکور و رتبه بندی گزینه ها بر اساس این مقدار Q_i

این مقدار برای هریک از آنها به صورت زیر تعریف می شود:

$$Q_i = v \left[\frac{S_i - S^*}{S^- - S^*} \right] + (1 - v) \left[\frac{R_i - R^*}{R^- - R^*} \right] \quad (6)$$

درجایی که $S^* = \text{Min}_i S_i$, $S^- = \text{Max}_i S_i$, $R^* = \text{Min}_i R_i$ و $R^- = \text{Max}_i R_i$ و v وزن استراتژی اکثریت موافق معیار یا حداکثر مطلوبیت گروهی است.

۳ پیشینه ی تحقیق

موسوی [۱] در تحقیق خود با عنوان "پیش پردازشی بر روش های تصمیم گیری چند معیاره با استفاده از ابزارهای داده کاوی" به لزوم کاربرد ابزارهای بهینه سازی چندهدفه در مسایل تصمیم گیری زندگی امروزی می پردازد. وی در ادامه به کاربرد روزافزون روش های تصمیم گیری چند معیاره در این راستا اشاره کرده، پیچیدگی این روش ها به دلیل تعداد رو به افزایش اجزای تشکیل دهنده، متغیرها، پارامترها، اعداد ثابت و اهداف موجود در این روش ها را مشکل اصلی آن ها می داند.

طلوع و همکاران [۳] در مقاله ای با عنوان "روشی جدید برای رتبه بندی الگوهای شناسایی شده در داده کاوی با استفاده از تحلیل پوششی داده ها" کاربرد روزافزون داده کاوی در تجارت را بیان می کنند. در استفاده از تکنیک های داده کاوی الگوهای مختلفی ممکن است به دست بیاید که تنها تعداد کمی از الگوها در اجرا استفاده می شود. ارزیابی و رتبه بندی جذابیت یا کارایی الگوهای وابستگی در داده کاوی نقش مهمی را ایفا می کند.

برخی از تحقیقات انجام شده در این شاخه به همراه رویکرد هر یک در جدول ۱ آورده شده است.

جدول ۱. برخی از تحقیقات انجام شده با روش های داده کاوی و تصمیم گیری چند معیاره [۶]، [۱۰-۱۸]

محقق	زمینه تحقیق
موسوی (۲۰۱۰)، هاله و همکاران (۲۰۱۲)	کاربر داده کاوی در پیش پردازش داده های ورودی روش های تصمیم گیری چند معیاره
مستروجنیس و همکاران (۲۰۰۹)، چویی و همکاران (۲۰۰۵)، موآتا و بریسن (۲۰۰۴)	تصمیم گیری چند معیاره نا رتبه ای در بهبود دقت الگوریتم های طبقه بندی داده کاوی
بابایی و همکاران (۱۳۹۰)	کاربرد روش های تصمیم گیری چند معیاره در پیش پردازش داده ها در الگوریتم های داده کاوی
فروغی (۲۰۱۱)، طلوع و نلچیکر (۲۰۱۱)، طلوع و همکاران (۲۰۰۹)	ارزیابی و رتبه بندی جذابیت یا کارایی الگوهای ارتباطی در داده کاوی با استفاده از روش های تصمیم گیری چند معیاره و تحلیل پوششی داده ها
راد و همکاران (۲۰۱۱)	مقایسه روش های تصمیم گیری چند معیاره با تکنیک های داده کاوی در رتبه بندی تعداد زیادی از گزینه ها
لیو و شیه (۲۰۰۵)	روش های تصمیم گیری چند معیاره با تکنیک های داده کاوی در سیستم های پیشنهاد محصول

۴ متدولوژی تحقیق

متدولوژی استفاده شده در این تحقیق شامل دو رویکرد داده کاوی و تصمیم گیری چند معیاره برای خوشه بندی و رتبه بندی حجم بالایی از رکوردها می باشد. چارچوب مدل ارایه شده در این تحقیق در شکل (۳) آورده شده است. مراحل اجرای مدل بر اساس استاندارد کریسپ در داده کاوی که پیش تر بیان شد، صورت گرفته است که در سه مرحله زیر پیاده سازی می گردد:

مرحله ۱- شناخت و انتخاب داده‌ها

مرحله ۲- پیش پردازش داده‌ها

مرحله ۳- مدل سازی و تجزیه و تحلیل داده‌ها و نتیجه گیری

۴-۱ شناخت و انتخاب داده‌ها

در مرحله نخست متغیرهای موجود در پایگاه داده شناسایی شده (این متغیرها در سال های گذشته توسط کارشناسان و خبرگان مرکز آمار ایران شناسایی شده است) و با توجه به نظر خبرگان و با استفاده از فرآیندهای تصمیم گیری چندمعیاره متغیرهای تأثیرگذار مشخص و تعریف می گردد.

پس از آماده شدن متغیرهای موردنظر و پایگاه داده، با استفاده از روش های تصمیم گیری چند معیاره مانند تحلیل سلسله مراتبی و با استفاده از نظر خبرگان و کارشناسان مربوط ۶ متغیر که بیش ترین تأثیر را بر نتیجه ی تحقیق حاضر دارد مشخص و اطلاعات خانوارها در مورد این ۶ متغیر به عنوان پایگاه داده جدید وارد فرآیند داده کاوی شده است.

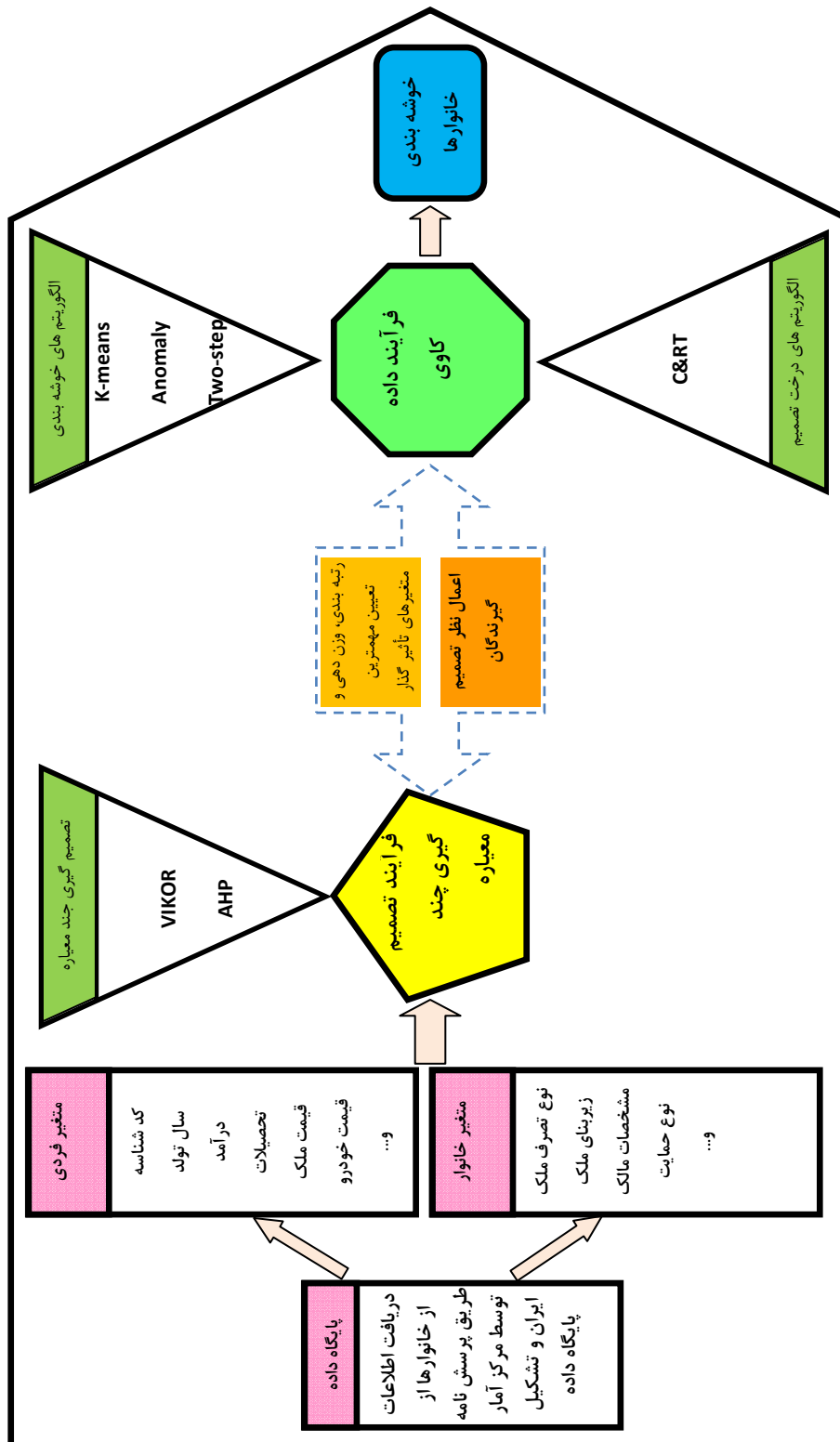
۴-۲ پیش پردازش داده‌ها

پیش پردازش شامل آماده سازی و نرمال سازی داده می شود. در بخش آماده سازی داده ها و برخورد با داده های گمشده، رکوردهای با بیش از ۵۰ درصد اطلاعات خالی حذف و در مورد سایر فیلدهای خالی، با توجه به متغیر مربوط تخصیص مقدار ثابت و یا استفاده از الگوریتم های پیش بینی کننده در داده کاوی در نظر گرفته شده است. در بخش دیگر آماده سازی داده ها که مربوط به برخورد با داده های غیر نرمال می باشد از الگوریتم آنومالی برای شناسایی داده های غیر نرمال استفاده شده و پس از شناسایی این داده ها رفتار مناسب در مورد آن ها صورت گرفته است. قبل از شروع اجرای مدل نیاز به نرمال سازی داده ها داریم. در تحقیق حاضر نرمال سازی داده ها با استفاده از روش Z استاندارد صورت گرفته است. این روش تمامی داده ها را به صورت اعدادی بین ۴ و -۴ تبدیل می کند.

۴-۳ مدل سازی

در این مرحله وزن متغیرها با توجه به نتایج به دست آمده از پرسش نامه ها و نظر خبرگان و با استفاده از روش تحلیل سلسله مراتبی محاسبه شده و بر روی داده های نرمال شده اعمال می گردد. سپس با توجه به شاخص انحراف استاندارد و شاخص سیلونت و تحلیل الگوریتم خوشه بندی دو مرحله ای تعداد خوشه ی مناسب برای خوشه بندی به روش کای میانگین مشخص شده و خوشه بندی با توجه به مقادیر وزن دار شده صورت می پذیرد. در آخر، خوشه های به دست آمده با استفاده از روش ویکور، رتبه بندی شده، تجزیه و تحلیل نهایی با توجه به خوشه و رتبه ی هر رکورد صورت می گیرد.

در ادامه در قالب یک مطالعه ی موردی به توضیح و تشریح مراحل انجام تحقیق پرداخته خواهد شد.



شکل ۳. چارچوب مدل تحقیق

۵ مطالعه‌ی موردی

مطالعه‌ی موردی این تحقیق مربوط به بانک اطلاعات یارانه‌های مرکز آمار ایران می‌باشد. پایگاه داده یکی از استان‌های کشور به عنوان جامعه‌ی آماری استفاده شده است.

۵-۱ مرحله‌ی شناخت و انتخاب داده‌ها

در مرحله‌ی نخست اطلاعات و فرم‌های بانک اطلاعات یارانه‌های خانوارهای یکی از استان‌های کشور در مرکز آمار ایران مورد بررسی قرار گرفته است. با توجه به اطلاعات موجود و طبق نظر خبرگان، ۲۴ متغیر جهت خوشه‌بندی خانوارها شناسایی شدند، شش متغیر تأثیرگذار و موردنیاز برای خوشه‌بندی با توجه به نظر خبرگان مرکز آمار ایران و با استفاده از پرسشنامه‌ی مقایسات زوجی، مشخص شده، اطلاعات هر خانوار در مورد این شش متغیر از بانک‌های اطلاعاتی موجود استخراج و به صورت فیلدهای جدیدی در جداول بانک اطلاعاتی تعریف و ذخیره‌سازی گردید. این شش متغیر شامل موارد زیر می‌شود:

- تحصیلات
- قیمت خودرو
- تعداد اعضا
- قیمت مسکن
- اقساط ماهیانه
- درآمد خانوار

۵-۲ مرحله‌ی پیش‌پردازش داده‌ها

۵-۲-۱ آماده‌سازی داده‌ها

در این مرحله ابتدا برخورد مناسب با داده‌های گمشده صورت گرفته است. در حالت اولیه تنها حدود ۱۷ درصد از فیلدهای اطلاعاتی پر می‌باشد که برای شروع داده کاوی این درصد باید به ۱۰۰ درصد افزایش یابد. برای این منظور از میان صدوپنجاه هزار خانوار استان، کلیه خانوارهایی که بیش از ۵۰ درصد اطلاعاتشان در بانک اطلاعاتی موجود نمی‌باشد از بانک اطلاعاتی کنار گذاشته می‌شود. با این کار درصد فیلدهای اطلاعاتی پر به ۴۹ درصد افزایش می‌یابد. در این مرحله تعداد خانوارهای باقیمانده در بانک اطلاعاتی یکصدودو هزار خانوار می‌باشد. در مرحله‌ی بعد برای فیلدهای خالی مربوط به متغیرهایی مانند قیمت خودرو و قیمت مسکن مقدار ثابت صفر در نظر گرفته شده است (چراکه این فیلدهای خالی مربوط به خانوارهایی بوده است که دارای مسکن یا خودروی شخصی نبوده‌اند).

در مورد فیلدهای خالی مربوط به متغیری مانند درآمد خانوار از الگوریتم‌های پیش‌بینی کننده در داده کاوی و مشخصاً الگوریتم کارت استفاده کرده‌ایم. در این بخش و قبل از استفاده از الگوریتم کارت، به دلیل حساسیت بالای این نوع الگوریتم‌ها به داده‌های پرت، ابتدا با استفاده از روش‌های آماری داده‌های پرت شناسایی شده (که در مورد تحقیق حاضر تعداد ۳۶ خانوار بوده) و موقتاً کنار گذاشته می‌شود و پس از اجرای الگوریتم پیش‌بینی کننده، دوباره به بانک داده‌ها اضافه می‌گردد. به این ترتیب مقادیر مربوط به درآمد خانوارهایی که درآمد خود را اعلام نکرده‌اند با توجه به سایر

اطلاعات خانوار در بانک اطلاعاتی، پیش بینی و در فیلد مربوطه قرار داده شده است. پس از این مرحله به ۱۰۰ درصد فیلدهای اطلاعاتی پر شده دست یافته ایم.

پس از تکمیل بانک اطلاعاتی نوبت به برخورد با داده های غیر نرمال می رسد. در پایگاه داده ی مورد مطالعه خانوارهایی دیده می شود که نسبت به سایرین غیرمتعارف یا غیر نرمال می باشند. در اینجا منظور ما داده های پرت و بافاصله زیاد نمی باشد. برای مثال فرض کنید یک خانوار اقساط ماهیانه ای برابر با بیست میلیون ریال پرداخت می کند در حالی که جمع در آمد خانوار چهار میلیون ریال می باشد. وجود چنین رکوردهایی در پایگاه داده موجب افزایش ضریب خطا در نتایج داده کاوی می گردد؛ لذا تشخیص و مدیریت این داده های غیرمتعارف در مرحله ی آماده سازی داده ها از اهمیت بسزایی برخوردار می باشد.

در مدل تحقیق حاضر پس از بررسی مطالعات انجام شده و آزمون روش های مختلف، استفاده از تکنیک و الگوریتم آنومالی که از تکنیک های خوشه بندی در داده کاوی به شمار می رود به عنوان یک روش کارا در تشخیص داده های غیرمتعارف استفاده شد. این تکنیک ضمن خوشه بندی داده ها، داده های غیرمتعارف را نیز در هر خوشه مشخص می کند. پس از اعمال الگوریتم آنومالی، رکوردهای با شاخص غیرمتعارفی بیش از ۳ به عنوان داده های غیرمتعارف در نظر گرفته می شود. اگر تعداد داده های غیرمتعارف زیاد باشد امکان حذف کل آنها وجود نخواهد داشت که در تحقیق حاضر به دلیل تعداد کم این داده ها (۲۰۰۰ رکورد) و پس از مشاوره با خبرگان مرکز آمار ایران داده های غیرمتعارف قبل از انجام خوشه بندی نهایی کنار گذاشته شده است. پس از مشخص شدن وضعیت داده های گمشده و داده های غیر نرمال پایگاه داده حاصل آماده ورود به مرحله ی بعد که همان فرآیند نرمال سازی است می باشد.

در این مرحله و برای بی مقیاس سازی داده ها روش های مختلف بی مقیاس سازی مورد آزمایش قرار گرفت و پس از مشاوره با خبرگان مرکز آمار روش استفاده از Z نرمال مناسب ترین روش برای این منظور تشخیص داده شد. در این روش از فرمول زیر برای بی مقیاس سازی فیلدهای مختلف استفاده می شود:

$$Z = \frac{X - @GLOBAL - MEAN(X)}{@GLOBAL_SDEV(X)} \quad (V)$$

در این فرمول X مقدار مربوط به هر فیلد است. @GLOBAL_MEAN میانگین مربوط به یک متغیر (ستون) خاص از جدول و @GLOBAL_SDEV مقدار انحراف از معیار یک متغیر خاص را برمی گرداند. لازم به توضیح است توابع گلوبال در نرم افزار کلمنتاین باید از قبل تعریف شود تا بتوان از آنها در مدل استفاده کرد که در مدل پروژه ی حاضر این کار صورت گرفته است.

۵-۳ مرحله‌ی مدل سازی و تجزیه و تحلیل داده‌ها و نتیجه گیری

در این مرحله نتایج نظر خبرگان در قالب پرسشنامه‌ی مقایسات زوجی و روش آقای ساعتی دریافت و پس از انجام محاسبات لازم نتایج به دست آمده به صورت وزن متغیرها بر روی داده‌ها اعمال شده است. به این صورت خوشه‌بندی با توجه به نظر خبرگان و اعمال اولویت‌های آن‌ها صورت می‌پذیرد. جدول مربوط به ماتریس نتایج به دست آمده از پرسش‌نامه‌ها و جدول وزن نهایی به دست آمده برای هر متغیر در ادامه آورده شده است.

جدول ۲. مقایسه زوجی معیارها

درآمد خانوار	تعداد اعضا	اقساط ماهیانه	قیمت خودرو	قیمت مسکن	تحصیلات
۰/۸۳۳۳۳۳۳۳۳	۲/۵	۱/۶۶۶۶۶۶۶۶۷	۰/۵۵۵۵۵۵۵۵۶	۰/۷۱۴۲۸۵۷۱۴	۱
۱/۱۶۶۶۶۶۶۶۷	۳/۵	۲/۳۳۳۳۳۳۳۳۳	۰/۷۷۷۷۷۷۷۷۸	۱	۱/۴
۱/۵	۴/۵	۳	۱	۰/۲۲۲۲۲۲۲۲۲	۱/۸
۰/۵	۱/۵	۱	۰/۳۳۳۳۳۳۳۳۳	۰/۴۲۸۵۷۱۴۲۹	۰/۶
۰/۳۳۳۳۳۳۳۳۳	۱	۰/۶۶۶۶۶۶۶۶۷	۰/۲۲۲۲۲۲۲۲۲	۰/۲۸۵۷۱۴۲۸۶	۰/۴
۱	۳	۲	۰/۶۶۶۶۶۶۶۶۷	۰/۸۵۷۱۴۲۸۵۷	۱/۲

جدول ۳. وزن معیارها

ردیف	متغیر	وزن محاسبه شده با روش تحلیل سلسله مراتبی
۱	تحصیلات	۰/۱۵۹۹۰۵۰۳۵
۲	قیمت مسکن	۰/۲۲۳۸۶۷۰۴۸
۳	قیمت خودرو	۰/۲۶۴۴۳۶۸۴۱
۴	اقساط ماهیانه	۰/۰۹۵۹۴۳۰۲۱
۵	تعداد اعضا	۰/۰۶۳۹۶۲۰۱۴
۶	درآمد خانوار	۰/۱۹۱۸۸۶۰۴۱

۵-۴ خوشه‌بندی

۵-۴-۱ تعیین تعداد خوشه‌ی مناسب

در این مرحله داده‌های وزن دار شده وارد فرآیند خوشه‌بندی می‌شود. برای اطمینان از تعداد خوشه‌ی مناسب در خوشه‌بندی یک فرایند سه مرحله‌ای در مدل صورت گرفته است:

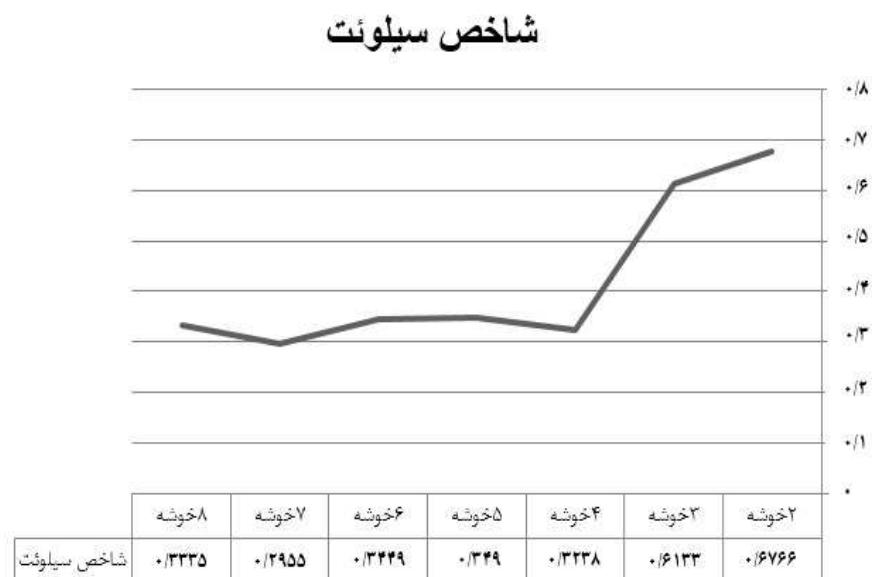
۵-۴-۱-۱ استفاده از الگوریتم دو مرحله‌ای

در ابتدا با استفاده از الگوریتم خوشه‌بندی دو مرحله‌ای خوشه‌بندی صورت گرفته است. این الگوریتم در طی انجام خوشه‌بندی، تعداد خوشه‌ی مناسب را تخمین می‌زند و خوشه‌بندی را انجام می‌دهد. در

این پژوهش این الگوریتم صرفاً جهت داشتن تخمینی از تعداد خوشه‌ی مناسب استفاده شده است که این تعداد برابر با دو خوشه‌ی به دست آمده؛ اما از آنجا که الگوریتم دو مرحله‌ای روشی مناسب برای تشخیص تعداد مناسب خوشه با ضریب اطمینان بالا نیست؛ لذا تعداد خوشه‌های پیشنهاد شده توسط این الگوریتم تنها به عنوان حدودی برای تعداد خوشه‌ی مناسب در نظر گرفته شده است.

۵-۴-۱-۲ شاخص سیلوئت

برای اطمینان از تعداد خوشه مناسب فرآیند خوشه‌بندی چندین بار و هر بار با تعداد خوشه‌ی مختلف اجرا شده و در نهایت خروجی مدل در هر بار به عنوان ورودی به نرم افزار متلب برده شده و در آنجا از شاخص سیلوئت جهت مقایسه‌ی تعداد خوشه‌ها و تشخیص تعداد خوشه‌ی مناسب استفاده شده است. نتایج این مقایسه در شکل ۴ آورده شده است.



شکل ۴. مقادیر شاخص سیلوئت با تعداد خوشه‌های مختلف

۵-۴-۱-۳ انحراف استاندارد

خوشه‌بندی با تعداد خوشه‌های ۲، ۳، ۴، ۵ و ۶ خوشه انجام شده و هر بار برای هر معیار و هر خوشه و نهایتاً هر خوشه‌بندی یک انحراف استاندارد به دست آمده است. نتایج محاسبات انحراف استاندارد خوشه‌بندی‌های مختلف در جدول ۴ آورده شده است.

با توجه به انحراف استاندارد به دست آمده در هر بار خوشه‌بندی مشاهده می‌شود تعداد خوشه‌ی سه مناسب‌ترین تعداد خوشه برای این مورد مطالعه می‌باشد. در نتیجه با توجه به تخمین به دست آمده از الگوریتم دو مرحله‌ای و مقایسه‌ی شاخص سیلوئت برای خوشه‌بندی با تعداد خوشه‌های مختلف و همچنین مقایسه‌ی

انحراف استاندارد در حالات مختلف خوشه‌بندی، تعداد خوشه‌ی سه مناسب‌ترین تعداد خوشه برای این مدل به‌دست آمده و خوشه‌بندی نهایی با در نظر گرفتن سه خوشه مبنای این تحقیق قرار گرفته است.

جدول ۴. انحراف استاندارد خوشه‌بندی‌های مختلف

انحراف استاندارد کل	انحراف استاندارد در هر خوشه						تعداد خوشه در نظر گرفته شده
	۶	۵	۴	۳	۲	۱	
۰/۱۵۵۱۶۶۶۶۷					۰/۱۴۲۵	۰/۱۶۷۸۳	۲
۰/۱۴۷۵۵۵۵۵۶				۰/۱۵۴۸۳	۰/۱۴۵۳	۰/۱۴۲۵	۳
۰/۳۲۲۶۲۵			۰/۹۲۲۸۳	۰/۱۲۲۱۶	۰/۱۲۳۵	۰/۱۲۲	۴
۰/۲۷۴۹۳۳۳۳۳		۰/۸۵۲۸۳	۰/۱۵۴۸۳	۰/۱۴۳۱۶	۰/۱۱۸۸۳	۰/۱۰۵	۵
۰/۳۲۰۷۲۲۲۲۲	۰/۱	۰/۱۶۳۶	۰/۸۵۲۸۳	۰/۵۹۹۱۶	۰/۱۰۴۶	۰/۱۰۴	۶

۵-۴-۲ خوشه‌بندی نهایی

در پایگاه داده یک ستون جدید اضافه می‌شود که مشخص‌کننده‌ی خوشه‌ی مربوط به هر خانوار می‌باشد. بخش کوچکی از پایگاه داده در جدول ۵ آورده شده است:

جدول ۵. پایگاه داده بعد از خوشه‌بندی

	W-C...	W-H...	W-VE...	W-E...	W-REV...	W-LOAN...	\$KM-K-Means
1	-0.093	0.167	-0.073	-0.234	-0.005	-0.024	cluster-1
2	0.147	-0.114	-0.073	0.060	-0.008	0.025	cluster-2
3	-0.053	-0.114	-0.073	-0.136	-0.010	-0.023	cluster-1
4	0.027	-0.069	-0.073	-0.057	-0.013	-0.036	cluster-1
5	-0.013	-0.114	-0.073	-0.087	-0.010	-0.021	cluster-1
6	-0.013	0.027	-0.073	-0.087	-0.005	-0.023	cluster-1
7	0.027	-0.058	-0.073	0.060	-0.002	-0.030	cluster-1
8	0.027	0.167	0.416	-0.175	0.009	0.026	cluster-1
9	0.107	0.055	-0.073	0.018	-0.010	-0.015	cluster-2
10	0.027	-0.102	-0.073	-0.057	-0.014	-0.036	cluster-1

نتایج به‌دست آمده بر اساس مقادیر هم‌مقیاس شده و وزن‌دار شده می‌باشد. برای تحلیل بهتر، میانگین نتایج به‌دست آمده برای هر خوشه از حالت نرمال و وزن‌دار خود خارج شده و در جدول ۶ نمایش داده شده است.

جدول ۶. میانگین معیارها در خوشه‌های به‌دست آمده

ردیف	متغیر	خوشه ۱	خوشه ۲	خوشه ۳
۱	تحصیلات	۱/۵۲۲۸۹۷	۱/۸۹۰۶۷۲	۲/۴۵۹۳۶۳
۲	قیمت مسکن	۱۷۵۹۴۸۳۲۱/۹	۲۳۴۷۷۶۰۹۸/۹	۲۳۸۳۴۱۴۱۸/۸
۳	قیمت خودرو	۱۲۰۲۲۳۶۴	۱۶۱۱۶۹۸۲	۲۴۵۱۰۹۴۹
۴	اقساط ماهیانه	۱۹۹۸۷۴۵	۲۲۹۱۱۵۴	۲۹۲۴۷۰۷
۵	تعداد اعضا	۳/۹۰۰۵۴۳	۶/۵۲۹۱۵۹	۲/۹۲۴۲
۶	درآمد خانوار	۲۹۱۱۲۳۱	۳۵۷۴۹۴۳	۴۶۳۶۸۸۲

۵-۴-۳ نام گذاری و تفسیر خوشه ها

با توجه به نتایج به دست آمده خوشه ها به صورت زیر نام گذاری می شود:

- I. خانوارهای با سطح درآمد و تحصیلات پایین
- II. خانوارهای پر جمعیت با سطح متوسط
- III. خانوارهای کم جمعیت با سطح تحصیلات و درآمد بالا

۵-۵ رتبه بندی خوشه ها با استفاده از روش راه حل توافقی و بهینه سازی چند معیاره (ویکور)

در این مرحله خوشه های به دست آمده از داده کاوی با توجه به میانگین مقادیر به دست آمده برای هر خوشه و در نظر گرفتن وزن های محاسبه شده برای متغیرها با استفاده از روش "راه حل توافقی و بهینه سازی چند معیاره" رتبه بندی شده است. متغیرها به عنوان معیارهای تصمیم گیری چند معیاره می باشد. با توجه به نتایج به دست آمده در جدول ۶ و پس از نرمال سازی ماتریس فوق، نتایج به دست آمده از روش راه حل توافقی و بهینه سازی چند معیاره مطابق جدول ۷ می باشد. همان طور که پیش تر بیان شد در این روش رتبه بندی نهایی بر اساس مقایسه ی مقادیر فاصله از راه حل ایده آل مثبت، فاصله از راه حل ایده آل منفی و مقدار شاخص ویکور به دست آمده است.

جدول ۷. رتبه بندی با روش ویکور

معیارها	تعداد اعضا	تحصیلات	قیمت مسکن	اقساط ماهیانه	درآمد خانوار	قیمت خودرو
وزن معیارها	۰/۰۶۳۹۶۲	۰/۱۵۹۹۰۵	۰/۲۲۳۸۶۷	۰/۰۹۵۹۴۳	۰/۱۹۱۸۸۶	۰/۲۶۴۴۳۶۸
خوشه ۱	۰/۴۷۸۶۹۳	۰/۴۴۰۶۸۳	۰/۴۶۵۴۷۱	۰/۴۷۳۷۷۱	۰/۴۴۵۲۲۳	۰/۳۷۹۲۱۸
گزینه ها خوشه ۲	۰/۸۰۱۲۸۹	۰/۵۴۷۱۰۷	۰/۶۲۱۱	۰/۵۴۳۰۸۲	۰/۵۴۶۷۲۶	۰/۵۰۸۳۷۴
خوشه ۳	۰/۳۵۸۸۷۲	۰/۷۱۱۶۶۹	۰/۶۳۰۵۳۲	۰/۶۹۳۲۵۶	۰/۷۰۹۱۳۱	۰/۷۷۳۱۴۳
f^*	۰/۳۵۸۸۷۲	۰/۷۱۱۶۶۹	۰/۶۳۰۵۳۲	۰/۴۷۳۷۷۱	۰/۷۰۹۱۳۱	۰/۷۷۳۱۴۳
f^-	۰/۸۰۱۲۸۹	۰/۴۴۰۶۸۳	۰/۴۶۵۴۷۱	۰/۶۹۳۲۵۶	۰/۴۴۵۲۲۳	۰/۳۷۹۲۱۸
رتبه بندی	مقدار (Qi)	مقدار (Vi)	فاصله از راه حل ایده آل منفی (Ri)	فاصله از راه حل ایده آل مثبت (Si)	گزینه ها	
۳	۱	۰/۵	۰/۲۶۴۴۳۷	۰/۸۵۷۴۱۸	خوشه ۱	
۲	۰/۵۰۸۰۱۶	۰/۵	۰/۱۷۷۷۳۶	۰/۴۹۹۹۷۸	خوشه ۲	
۱	۰	۰/۵	۰/۰۹۵۹۴۳	۰/۰۹۵۹۴۳	خوشه ۳	

این رتبه‌بندی بر مبنای سطح رفاه خانوارها صورت گرفته است و می‌تواند در تصمیم‌گیری‌های آتی مرکز آمار ایران و تدوین استراتژی‌های این مرکز در تخصیص منابع کاربرد داشته باشد. با این رتبه‌بندی درواقع برای هر خانوار با توجه به خوشه‌ی مربوط یک رتبه در نظر گرفته شده است.

۵-۶ نام‌گذاری و توصیف خوشه‌ها

با توجه به نتایج به‌دست آمده، خوشه‌ها به‌صورت زیر نام‌گذاری و توصیف می‌شود:

۵-۶-۱ خانوارهای با سطح درآمد و تحصیلات پایین

این خوشه با ۵۶٪ خانوارهای استان بیش‌ترین درصد خانوارهای استان را به خود اختصاص داده است. خانوارهای این خوشه پایین‌ترین سطح درآمد را در بین کل خانوارهای استان مورد مطالعه داشته و هم‌زمان با سطح درآمد پایین سطح تحصیلات نیز در این خانوارها بسیار پایین می‌باشد. به‌طوری‌که اکثر اعضای این خانوارها را افراد کم‌سواد تشکیل می‌دهد.

به نظر می‌رسد استفاده از معیار شهرنشینی و روستانشینی خانوارها در تحلیل این خوشه موردنیاز باشد چراکه خانوارهایی با سطح درآمد و تحصیلات پایین و جمعیت متوسط در روستا دارای رفاه نسبی بوده درحالی‌که خانواری با این شرایط در محیط شهر از اقشار ضعیف جامعه محسوب می‌شود.

۵-۶-۲ خانوارهای پرجمعیت با سطح متوسط

این خوشه ۲۴٪ جمعیت کل استان را شامل می‌شود. مشخصه‌ی بارز خانوارهای این خوشه پرجمعیت بودنشان است. به‌طوری‌که میانگین جمعیت خانوارها در این خوشه ۶/۵ نفر می‌باشد. خانوارهای این خوشه در سایر معیارها وضعیتی متوسط دارند؛ اما تعداد زیاد افراد خانوارهای این خوشه وضعیت رفاهی افراد این خانوارها را به‌شدت تحت تأثیر قرار می‌دهد. جمعیت زیاد این خانوارها در حالی است که دو خوشه‌ی دیگر دارای میانگین جمعیت ۳ و ۴ نفر در هر خانوار می‌باشد.

۵-۶-۳ خانوارهای کم‌جمعیت با سطح تحصیلات و درآمد بالا

خانوارهای این خوشه با بالاترین سطح تحصیلات در استان دارای کم‌ترین میانگین جمعیت خانوار (۳) و همچنین بالاترین سطح درآمد می‌باشند. با توجه به نتایج رتبه‌بندی از نظر وضعیت رفاهی خانوارها و بر اساس معیارها و وزن‌های ذکرشده به دست آمدن بالاترین رتبه برای این خوشه‌ی طبیعی و مطابق با نتایج خوشه‌بندی می‌باشد.

همچنین این خوشه به دلیل سطح بالای تحصیلات خانوارها دارای مطمئن‌ترین داده‌ها در تحقیق می‌باشد و در مدل‌های پیش‌بینی می‌توان از این داده‌ها به‌عنوان داده‌های آزمایشی استفاده نمود.

۶ نتیجه گیری

۶-۱ امکان کاربردی شدن هر چه بیش تر مدل های داده کاوی توسط مدیران

مدل ارایه شده در پژوهش حاضر با وارد کردن نظرات مدیران در فرآیند داده کاوی تأثیر تصمیمات مدیران بر نتایج داده کاوی را پررنگ تر می سازد و این امر می تواند موجب رغبت هر چه بیش تر مدیران در استفاده از تکنیک های داده کاوی گردد. این مدل و ایده استفاده از وزن و رتبه های به دست آمده با روش های تصمیم گیری چند معیاره در داده کاوی قابل اجرا در سایر روش های داده کاوی از جمله مدل های درخت تصمیم و تکنیک های پیش بینی می باشد.

۶-۲ رتبه بندی خوشه ها

رتبه بندی خوشه ها خود یک رویکرد نوین در این زمینه است که می تواند راهنمای مدیران در تصمیم گیری ها باشد. این رتبه بندی می تواند از دیدگاه های مختلف و با روش های مختلف صورت بگیرد. برای مثال در تخصیص منابع می توان از روش های وزن دهی استفاده کرده، تخصیص منابع را بر اساس وزن های به دست آمده انجام داد.

۶-۳ مزیت کاربرد این مدل در مرکز آمار ایران

استفاده از این مدل به عنوان یک مدل ترکیبی موجب افزایش دقت تعیین خوشه ها خواهد شد. همچنین رتبه بندی انجام شده که ورودی خود را از خروجی خوشه بندی انجام شده گرفته است باعث مشخص شدن سطح رفاه خانوارها در هر خوشه می شود و این امر از بروز خطاهای احتمالی در تعیین سطح رفاه خانوارها تا حد زیادی جلوگیری خواهد کرد؛ البته دسترسی به اطلاعات خوشه بندی های قبلی امکان پذیر نبوده، در صورت استفاده از این مدل در خوشه بندی های آتی امکان مقایسه وجود خواهد داشت.

منابع

- [۲] کریمی شیرازی، ح.، مدیری، م. و فرچورخانا پشتهانی، ق.، (۱۳۹۳). یک مدل MCDM جدید ترکیبی از VIKOR، ANP و DEMATEL برای اولویت بندی کاربردهای فناوری نانو در بخش صنایع غذایی. مجله تحقیق در عملیات و کاربردهای آن، ۹(۲)، ۳۳-۱۳.
- [۶] غضنفری، م.، عزیزاده، س. و تیمورپور، ب.، (۱۳۸۷). داده کاوی و کشف دانش (نسخه چاپ اول). تهران، دانشگاه علم و صنعت ایران.
- [۷] آذر، ع.، سبط، و.، (۱۳۸۹). طراحی مدل انتخاب نیروی انسانی با رویکرد داده کاوی. نشریه مدیریت فناوری اطلاعات، ۲(۴)، ۳-۲۲.
- [۸] اصغر پور، م. ج.، (۱۳۸۷). تصمیم گیری های چند معیاره. موسسه انتشارات و چاپ دانشگاه تهران.
- [۱۴] بابایی، م.، صفار یزدی، ز. و سرایی، م. ح.، (۱۳۹۰). معرفی و مقایسه روش های پیش پردازش داده برای کاربردهای مختلف داده کاوی. تکنولوژی و علوم دانشگاه اصفهان.

- [1] Mosavi, A.,(2010). Multiple Criteria Decision Making Preprocessing Using Data Mining Tools. International Journal of Computer Science Issues, 7(2): 26-34.
- [3] Toloo, M., Sohrabi, B., Nalchigar, S., (2009). A new method for ranking discovered rules from data mining by DEA. Expert Systems with Applications. 36: 8503-8508.
- [4] Coenen, F., (2011). Data mining: past, present and future. The Knowledge Engineering Review. 26(1): 25-29.
- [5] Fayyad, U., Piatesky-Shapiro, G., Smyth, P., (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of The ACM. 37: 27-34.
- [9] Jingzhu, W., Xiangy, L., (2009). The Multiple Attributed Decision-Making VIKOR Method and Its Application. IEEE.
- [10] Haleh, H., Ghaffari, A., Meshki, A. K ., (2012). A Combined Model of MCDM and Data Mining for Determining Question Weights in Scientific Exams .Applied Mathematical Sciences. 6: 173-196.
- [11] Mastrogiannis, N., Boutsinasa, B., Giannikos, I.,(2009). A method for improving the accuracy of data mining classification algorithms. Computers&Operations Research. 36: 2829-2839.
- [12] Choi, D. H., Ahn, B. S., Kim, S. H.,(2005). Prioritization of association rules in data mining: Multiple criteria decision approach. Expert Systems with Applications. 29: 867-878.
- [13] Muata, K., Bryson, O., (2004). Evaluation of decision trees: a multi-criteria approach. Computers&Operations Research. 34: 1933-1945.
- [15] Foroughi, A., (2011). A note on "A new method for ranking discovered rules from data mining by DEA," and a full ranking approach. Expert Systems with Applications. 38: 12913-12916.
- [16] Toloo, M., Nalchigar, S., (2011). A new DEA method for supplier selection in presence of both cardinal and ordinal data . Expert Systems with Applications. 38: 14726-14731.
- [17] Rad, A., Naderi, B., Soltani, M., (2011). Clustering and ranking university majors using data mining and AHP algorithms: A case study in Iran. Expert Systems with Applications. 38: 755-763.
- [18] Liu, D., Shih, Y., (2005). Integrating AHP and data mining for product recommendation based on customer lifetime value. Information&Management. 42: 387-400.