
بسم الله الرحمن الرحيم

مرکز آموزشهای الکترونیکی

گروه مهندسی کامپیوتر

پایان نامه کارشناسی ارشد

کاربرد داده کاوی در رده بندی تخلفات ساختمانی و تحلیل انگیزه های تخلف

داود مزجی

استاد راهنما :

دکتر حمید حسن پور

بهمن ماه ۱۳۹۴

دانشگاه شاهرود
مدیریت تحصیلات تکمیلی
فرم شماره (۶)

باسمه تعالی

شماره: ۱۰۳/۱۹۰۹الف
تاریخ: ۹۴/۱۲/۲۴
ویرایش:

فرم صورت جلسه دفاع از پایان نامه تحصیلی دوره کارشناسی ارشد

با تأییدات خداوند متعال و با استعانت از حضرت ولی عصر (عج) نتیجه ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم / آقای داود مزجی به شماره دانشجویی ۹۱۲۶۶۶۴ رشته مهندسی کامپیوتر گرایش هوش مصنوعی تحت عنوان کاربرد داده کاوی در دسته بندی تخلفات ساختمانی و تحلیل انگیزه های تخلف

که در تاریخ ۹۴/۱۱/۲۶ با حضور هیأت محترم داوران در دانشگاه شاهرود برگزار گردید به شرح ذیل اعلام می گردد:

قبول (با درجه : بسیار خوب امتیاز ۱۸/۵) ☑ دفاع مجدد ☐ مردود ☐

۱ـ عالی (۲۰ ـ ۱۹)
۲ـ بسیار خوب (۱۸/۹۹ ـ ۱۸)
۳ـ خوب (۱۷/۹۹ ـ ۱۶)
۴ـ قابل قبول (۱۵/۹۹ ـ ۱۴)
۵- نمره کمتر از ۱۴ غیر قابل قبول

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استادراهنما	حمید حسن پور	استاد	
۲ـ استاد مشاور	مرتضی زاهدی	استادیار	
۳ـ نماینده شورای تحصیلات تکمیلی	محسن فرهادی	مربی	
۴ـ استاد ممتحن	علی اکبر پویان	استادیار	
۵ ـ استاد ممتحن	هدی مشایخی	استادیار	

رئیس دانشکده: امضاء

وزارت علوم، تحقیقات و فناوری
آموزشهای عالی

چکیده

بخش عمده ای از مسائل و ناهنجاری های شهرها را می توان به تخلفات کَمّی ساختمانی نسبت داد، مطالعه و بررسی دلایل و انگیزه های تخلف، و کشف ارتباط بین پارامترها و متغیرهای موثر در وقوع تخلف با آن، در هدایت شهرها به سمت توسعه نقش بسزایی خواهد داشت.

در این پژوهش کشف الگوی پنهان از بانک اطلاعات ساخت و ساز شهری برای استخراج روابط بین این متغیرها و تخلفات ساختمانی مد نظر قرار گرفته است. چگونگی گردآوری داده ها و ترکیب ساختار آنها از مهمترین و زمانبرترین بخش های پژوهش بشمار آمده و مهمترین اولویت در گردآوری داده ها تلاش برای حفظ ارزش آنها بوده است.

در این مسیر مطالعهٔ روشها و رویکردهای سایر پژوهشگران در زمینه کار با داده های ترکیبی ، اهمیت انتخاب و استخراج ویژگی، آماده سازی و پیش پردازش شامل تکنیکهای پاکسازی، یکپارچه سازی، کاهش و تبدیل داده صورت پذیرفته و ضمن ارزیابی الگوریتم K-Means بر داده ها، اهمیت معیارهای شباهت و شاخصهای اعتبار سنجی خوشه ها و نیز مقایسه خروجی با خروجی الگوریتمهای فراابتکاری خودکار PSO و GA مورد بررسی قرار گرفته است.

نتایج پژوهش با در نظر گرفتن صحت انتخاب تعداد خوشه و ارزیابی اثر تعداد ویژگیها در نتیجه، بصورت جداول مقایسه ای و با نمایش شاخص های مرکزی هر ویژگی در هر خوشه سازمان یافته و به صورت ارائه گزاره های منطقی آمده است. در پایان برخی نتایج از جمله ارتباط قیمت منطقه ای املاک، مساحت آنها یا ساخت و ساز در حریم با تخلفات ساختمانی ارائه گردیده است.

واژگان کلیدی : داده کاوی، داده ترکیبی، تخلفات ساختمانی

فهرست مطالب

فهرست شکل ها :

فهرست جداول :

فصل اول

کلیات پژوهش

۱-۱- مقدمه

تدوین ضوابط و مقررات شهرسازی که ماحصل آن الگوهای رایج در ساخت و سازهای شهری است با این هدف که شهرها بر اساس اصول و قوانین صحیح و اندیشیده شده و با توجه به قابلیت های اقتصادی، اجتماعی، فرهنگی و منطقه ای رشد و توسعه یابند، صورت پذیرفته است. این درحالی است که بر اساس مطالعات انجام شده در طول ۵۰ سال گذشته ۲۰ درصد از اهداف طرح های جامع شهری تهیه شده در کشور، محقق شده است (۱).

عدول از ضوابط طرح های جامع و تفصیلی شهرها منتج به تخلفاتی است که بخش عمده آن به تخلفات ساختمانی مالکان مربوط می گردد. این تخلفات به دو دسته کمی و کیفی تقسیم می شوند.

- تخلفات کیفی با جان مردم در ارتباط است مانند اینکه مالک ، ضوابط آتش نشانی را در زمان ساخت ملک خود رعایت نکند، یا اینکه پلهٔ فرار برای ساختمان های بلند مرتبه در نظر گرفته نشود و یا اینکه در زمان ساخت، در ساختمان درز انقطاع که در زمان وقوع زلزله و تخریب ساختمان ها بسیار تاثیرگذار است، ایجاد نشود و مهمتر از همه عدم وجود استحکام لازم و کافی و عدم وجود استانداردهای لازم در شهرسازی.

  بدیهی است که در این گونه موارد با عدم صدور گواهی پایان کار برای ساختمان های متخلف و غیر مجاز، با این تخلفات برخورد شود.

- اما بخش عمده ای از مسائل و ناهنجاری های شهرها را می توان به تخلفات کَمّی ساختمانی نسبت داد که پس از رشد مهاجرت به شهر و رونق ساخت و سازها، در شهر ها شدت یافته است. برخورد با این تخلفات از طریق اقداماتی چون اخذ عوارض یا جریمه، اعاده به وضع سابق، تامین پارکینگ،

اصلاح پخی و یا رفع تعرض صورت می پذیرد. اما متداول ترین آن اخذ عوارض و جریمه که به رواج فروش تراکم و ضابطه نیز مشهور است در غالب آراء مشاهده می گردد.

۱-۱-۱- ضرورت مطالعه تخلفات ساختمانی و مقابله با آنها

تبعات بروز تخلفات ساختمانی در صورت برخورد با این تخلفات ، نارضایتی ، مقاومت ، عدم همکاری و مشارکت و همراهی با مدیریت شهری بوده و در صورت عدم برخورد با آنها ، شهری با توسعه هدایت نشده و ساخت و سازهای لجام گسیخته خواهد بود که علاوه بر نارضایتی مجاورین و همسایگان ممکن است امنیت و آسایش را در حوزه ای وسیعتر تحت تاثیر خود قرار دهد. بعنوان مثال عدم رعایت ضوابط مربوط به تامین پارکینگ مشکلات ترافیکی معابر اصلی و پر تردد را علاوه بر ساکنین املاک متوجه عموم شهروندان خواهد ساخت. چه آنکه در زمان بروز بحران خود گره ای بر سایر گره ها و چالشی برای مدیریت بحران خواهد بود. از اینرو مطالعه و بررسی دلایل و انگیزه های تخلف ، تحلیل زمانی و مکانی آن ، تحلیل رفتار شهروندان در قبال برخی قوانین و رویکردها ، تحلیل حوادث، تحلیل سلایق شهروندان ، تحلیل رضایتمندی شهروندان، تاثیر عوامل محیطی و فرهنگی بر نوع تخلفات، بهره گیری از اطلاعات دسته بندی شده در ارائه خدمات متناسب و غیره در هدایت شهر به سمت توسعه ، نقش بسزایی خواهد داشت.

اگر چه در گذشته در خصوص انگیزه های تخلف و بررسی و تحلیل آن (۲) و همچنین ارائه تحلیل آماری از انواع تخلفات و انگیزه های مربوط به آنها (۳) ، موانع اجرایی ضوابط شهرسازی و ارائه راهکارهای مناسب در جهت جلوگیری از تخلفات ساختمانی (۴) بررسی تاثیر تصمیمات کمیسیون ماده ۱۰۰ شهرداری در کنترل تخلفات ساختمانی (۵)، ریشه یابی علل تخلفات ساختمانی و دلایل تخلفات از جمله عدم آگاهی شهروندان از قوانین و مقررات شهرسازی، عدم اطلاع رسانی صحیح شهرداری ها در زمان صدور پروانه ساخت، وضعیت اقتصادی شهروندان ، عدم تعیین وضعیت اراضی با کاربری های عمومی از طرف متولیان کاربری، سهل انگاری و عدم نظارت صحیح بر ساخت و سازها توسط شهرداری ها (ضعف سیستم کنترل و نظارت) ، ضعف ارگان های دولتی در محافظت از

اراضی و املاک تحت مالکیت دولت (۶) و عدم وجود کنترل زمین شهری و ناسازگاری ضوابط با واقعیات جامعه، وجود قوانین مبهم و غیرشفاف، بلاتکلیفی اراضی قولنامه ای، نگاه شهرداری ها به جریمه ها (به جای تخریب یا اعاده به وضع سابق) به عنوان یک منبع درآمد (۷) و موضوعات مشابه، تحقیقات خوبی در شهرهای مختلف کشور از جمله تهران ، مشهد ،تبریز، یزد و غیره صورت پذیرفته است. لیکن بررسی بر اساس کشف دانش یا الگوی پنهان از بانکهای اطلاعاتی شهری سابقه چندانی نداشته و موضوعی است که پرداختن به آن مفید به نظر می رسد.

۱-۱-۲- سایر عوامل موثر بر تخلفات ساختمانی

همچنین در نگاهی دیگر به پژوهش های صورت پذیرفته در خصوص تخلفات ساختمانی و از دید نگارنده موارد متعدد دیگری از جمله موراد ذیل می تواند در وقوع تخلفات ساختمانی موثر باشند.

۱-۱-۲-۱- تعرفه عوارض محلی

با عنایت به اینکه شوراهای اسلامی شهرها وظیفه تدوین قوانین محلی (از جمله تعرفه عوارض محلی شهرداریها که بخش قابل توجهی از آن به ساخت و سازها معطوف می گردد) را بر عهده دارند و در این قوانین با هدف « تضعیف منافع مالی تخلفات ساختمانی» اقدام به تعیین جرایم مالی می نمایند ، در فرآیند ساخت و ساز در شهرها موثر هستند. عامل موثر در تعرفه عوارض مربوط به ساخت و ساز ، قیمت منطقه ای یا ارزش معاملاتی زمین (موضوع ماده ۶۴ قانون مالیاتهای مستقیم[1]) می باشد.

---

۱ - تعیین ارزش معاملاتی املاک بر عهده کمیسیون تقویم املاک می‌باشد. کمیسیون مزبور موظف است ارزش معاملاتی موضوع این قانون را در سال اول معادل دودرصد(٪۲) میانگین قیمتهای روز منطقه با لحاظ ملاکهای زیر تعیین کند. این شاخص هرسال به‌میزان دوواحددرصد افزایش می‌یابد تا زمانی که ارزش معاملاتی هر منطقه به بیست‌درصد(٪۲۰) میانگین قیمتهای روز املاک برسد.

الف ـ قیمت ساختمان با توجه به مصالح(اسکلت فلزی یا بتون آرمه یا اسکلت بتونی و سوله و غیره) و قدمت و تراکم و طریقه استفاده از آن(مسکونی، تجاری، اداری، آموزشی، بهداشتی، خدماتی و غیره) و نوع مالکیت

۱-۱-۲-۲- منطقه ای که ملک در آن واقع شده

بافت جمعیتی مناطق مختلف شهر که پیرو مسائلی چون فرسودگی بافت ، مراکز تفریحی و گردشگری، مراکز خرید یا اداری و غیره می باشد می تواند از عوامل موثر در نوع تخلفات ساختمانی به شمار آید. انعکاس این عامل در بانک اطلاعاتی ، قیمت منطقه ای املاک می باشد که هر ساله ضوابط مربوط به آن توسط اداره دارایی ابلاغ می گردد.

۱-۱-۲-۳- طرح تفصیلی شهر

طرح تفصیلی، عبارت از طرحی است که بر اساس معیارها و ضوابط کلی طرح جامع شهر[۲]، نحوه استفاده از زمین های شهری در سطح محلات مختلف شهر و موقعیت و مساحت دقیق زمین برای هر یک از آن ها و وضع دقیق و تفصیلی شبکه عبور و مرور و میزان تراکم جمعیت و تراکم ساختمانی در واحدهای شهری و اولویت های مربوط به مناطق به سازی و نوسازی و توسعه و حل مشکلات شهری و موقعیت کلیه عوامل مختلف شهری در آن تعیین می شود و نقشه ها و مشخصات مربوط به مالکیت بر اساس مدارک ثبتی تهیه و تنظیم می گردد.

متاسفانه اعمال برخی شرایط غیر منصفانه در اجرای این طرح توسط مشاور، می تواند موجب افزایش یا کاهش قیمت املاک شهروندان گردد. که قانونگذار در یک راه کار و برای رفع این

ب ـ قیمت اراضی باتوجه به نوع کاربری و موقعیت جغرافیایی از لحاظ تجاری، صنعتی، مسکونی، آموزشی، اداری و کشاورزی این کمیسیون متشکل از پنج عضو است که در تهران از نمایندگان سازمان امور مالیاتی کشور، وزارتخانه‌های راه و شهرسازی و جهادکشاورزی، سازمان ثبت اسناد و املاک کشور و شورای اسلامی شهر و در سایر شهرها از مدیران کل یا رؤسای ادارات امور مالیاتی، راه و شهرسازی، جهاد کشاورزی و ثبت اسناد و املاک و یا نمایندگان آنها و نماینده شورای اسلامی شهر تشکیل می‌شود. کمیسیون مذکور هر سال یک‌بار ارزش معاملاتی املاک را به تفکیک عرصه و اعیان تعیین می‌کند.

۲- طرح جامع، طرح بلندمدتی است که در آن نحوه استفاده از اراضی و منطقه بندی مربوط به حوزه های مسکونی، صنعتی، بازرگانی، اداری و کشاورزی و تاسیسات و تجهیزات و تسهیلات عمومی مناطق نوسازی، به سازی و اولویت های مربوط به آن ها تعیین شده و ضوابط و مقررات مربوط به کلیه مواردو همچنین ضوابط مربوط به حفظ بنا و نماهای تاریخی و مناظر طبیعی، تهیه و تنظیم می گردد و برحسب ضرورت، قابل تجدیدنظر خواهد بود. (طبق قانون تغییر نام وزارت آبادانی و مسکن به وزارت مسکن و شهرسازی و تعیین وظایف آن)

مشکلات، با ارجاع پرونده ها به کمیسیون ماده ۵ و تغییر کاربری ها توسط این کمیسیون اصلاح طرح را پیش بینی نموده است. لذا میزان ارجاعات به کمیسیون ماده ۵ و اصلاحات این کمیسیون می تواند شاخصی برای ارزیابی عملکرد مشاور و طرح تفصیلی باشد.

۱-۱-۲-۴- عملکرد کمیسون ماده ۱۰۰

این ویژگی توسط مقایسه فیلد نوع رای با نوع تخلف قابل ارزیابی است. (بدلیل تفاوت در حل مساله در این پژوهش بررسی نمی شود)

۱-۲- طرح مسئله

موضوعات فوق الذکر از جنبه های گوناگون قابل بررسی و پیگیری است. مثلاً شیوه های مبتنی بر اخذ دانش از یک نمونه آماری مناسب و بر اساس مدلهای استاندارد رایج و تحلیل نتایج (روش بکار گرفته شده در عموم موارد تحقیقی در این زمینه) و یا روش استخراج ویژگیهای موثر از بانک اطلاعات شهری و دسته بندی و تحلیل نتایج آن (که در این پژوهش به آن خواهیم پرداخت).

بعبارت دیگر این پژوهش پاسخ پرسشهای ذیل را فراهم خواهد نمود:

- آیا می توان از طریق بررسی اطلاعات املاک دارای تخلف ساخت و ساز در شهر به روابطی منطقی برسیم که راهنمای ما در رسیدن به دلایل تخلف باشد؟
- آیا تمایل شهروندان منطقه ای از شهر (و یا حتی تمام شهر) به رفتارهایی که مغایر برخی ضوابط شهری است، معنادار است؟
- آیا اصلاً رده بندی املاک دارای تخلفات ساختمانی می تواند مبین شاخص هایی برای هر دسته باشد؟ و آیا این شاخص ها برای تصمیم گیری های آینده موثر است؟
- آیا مشخصات املاک از جمله مساحت یا منطقه ای که ملک درآن واقع شده می تواند بیانگر وقوع نوع خاصی از تخلف در ساختمان آن ملک باشد؟

در پاسخ به این سوالات نیز سوالات فرعی زیر مطرح خواهد شد:

- ویژگیهای بیان کننده مسئله چگونه انتخاب خواهند شد؟
- مزیت ویژگیهای انتخاب شده بر سایر ویژگیها چیست؟

بهره گیری از داده کاوی بانک اطلاعات شهری در رده بندی انواع تخلفات ساختمانی با توجه به موقعیت مکانی (مختصات و شرایط آن) ، کاربری ، طرح تفصیلی (مثلاً تراکم پایه و مجاز) و سایر مشخصات مورد نظر (با توجه به محتویات بانک های مربوط به مطالعه موردی) ایده ای است که در این پژوهش به آن پرداخته می شود. بعبارت دیگر این پژوهش در پی یافتن ارتباط بین انواع تخلفات صورت پذیرفته و عوامل موثر در وقوع تخلف می باشد. بدیهی است که ارزیابی ویژگیهای هم گروه (موجود در یک خوشه) یا بعبارت دیگر یافتن ارتباط بین ویژگیها و تخلفات میتواند در جهت اصلاح رویه ها و قوانین محلی ، طرح های جامع و تفصیلی و همچنین کمک به نظارت بر ساخت و سازها مفید باشد.

بصورت موردی، یافتن ارتباط بین قیمت منطقه ای و انواع تخلفات ، ارتباط بین متراژ زمین و تخلف ، ارتباط بین کسری پارکینگ و زیر بنا و یا ارتباط بین مازاد تراکم و متراژ معبر وقوع می تواند از نتایج ارزشمند این پژوهش باشد. بعنوان مثال اینکه زمینهای دارای قیمت منطقه ای A عمدتاً دارای تخلف B هستند. و نتیجه این ارتباط می تواند منتج به تصمیم گیری راجع به ضوابط ، تعرفه محلی یا افزایش کنترل بر ساخت و ساز شود.

در بهترین وضعیت (یعنی وجود و کشف الگوهای پنهان)، از دیگر نتایج این پژوهش می تواند رده بندی داده های جدید باشد بعبارت دیگر پیش بینی نوع تخلف یا تخلفات، قبل از مراجعه حضوری و بر اساس مشخصات ورودی یک رکورد جدید می تواند از نتایج پژوهش باشد.

۱-۳- داده کاوی [۳]

[۳] - DataMining

آرشیوهای اطلاعاتی، به دلیل حجم بسیار زیاد، غالبا به مقبره های اطلاعات تبدیل می شوند و از قابلیتهای بالقوه اطلاعات ذخیره شده استفاده نمی شود. داده کاوی به بهره گیری از ابزارهای تجزیه و تحلیل این داده‌ها به منظور کشف الگوها و روابط معتبری که تا کنون ناشناخته بوده‌اند اطلاق می‌شود. این ابزارها ممکن است مدلهای آماری، الگوریتم‌های ریاضی و روش‌های یاد گیرنده[۴] باشند که کار خود را به صورت خودکار و بر اساس تجربه‌ای که از طریق الگوریتم های یادگیر به دست می‌آورند بهبود می‌بخشند.

در بین روشهای مختلف طبقه بندی داده، دو روش مهم در داده کاوی عبارتند از:

- دسته بندی یا کلاسبندی[۵] : فرآیند پیدا کردن مدلی برای پیش بینی رکوردهایی که برچسب دسته آنها از قبل شناخته شده هست.
- خوشه بندی[۶] : فرآیند پیدا کردن مدلی برای پیش بینی رکوردهایی که برچسب دسته آنها از قبل شناخته شده نیست. و رده بندی رکوردها یا اشیاء به نحوی در آن صورت می پذیرد که اعضای موجود در یک خوشه بیشترین شباهت را به یکدیگر و کمترین شباهت را به اعضای خوشه‌های دیگر داشته باشند.

در مساله داده کاوی دو مقدمه مهم است یکی فرمول واضحی از مساله که قابل حل باشد و دیگری دسترسی به داده متناسب. بعضی از ناظران داده کاوی را مرحله‌ای در روند کشف دانش در پایگاه داده‌ها[۷] می‌دانند.

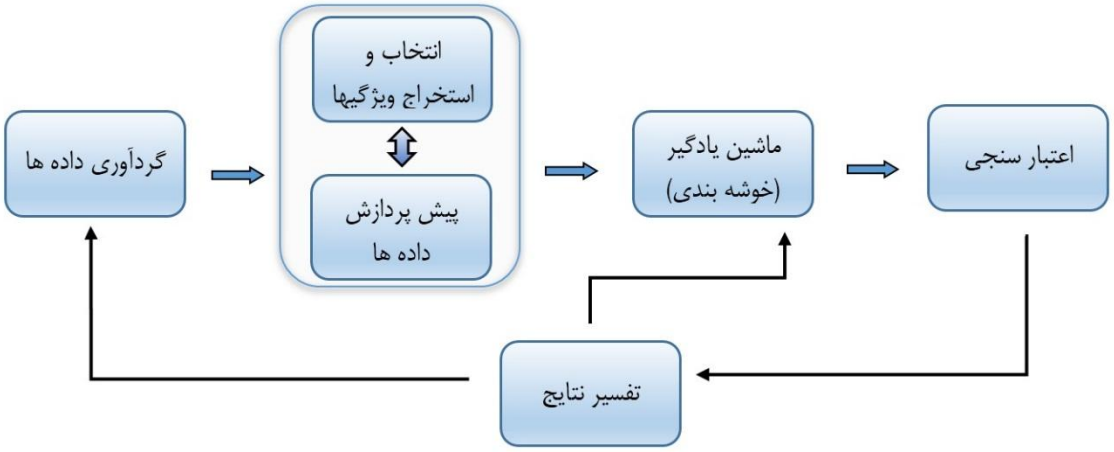
۴- Machine Learning Methods

۵ - Classification

۶ - Clustering

۷ - Knowledge Discovery in Database (KDD)

۱-۴- گام های پژوهش



شکل ۱،۱ : گامهای پژوهش

۱-۴-۱- گردآوری داده ها

۱-۴-۱-۱- مطالعه موردی

در شهر شاهرود اکثر ساخت و سازها کم و بیش دارای تخلفات خصوصاً تخلفات کَمّی هستند بطوریکه از میان ۷۵۰۰ پروانه صادره در بازه زمان سالهای ۱۳۸۷ تا ۱۳۹۴، بیش از ۸۰٪ گواهیهای پایان کار با چنین تخلفاتی ثبت شده اند. همچنین با عنایت به بومی بودن نگارنده و آگاهی نسبی از ساختارهای اداری ، فرهنگی و اجتماعی و از همه مهمتر امکان دسترسی به بانک اطلاعات املاک شهرداری شاهرود، این شهر گزینه مناسبی برای پژوهش به نظر می رسد.

این داده ها مشتمل بر بیش از شانزده هزار رکورد تخلفات ساختمانی (ارجاعی به کمیسیون ماده ۱۰۰) بوده و دلایل تفاوت بین تعداد پروانه های صادره و تعداد رکورد های تخلف به شرح ذیل می باشد :

- برخی تخلفات مربوط به پرونده های قبل از دوره مکانیزاسیون شهرداری می باشد که با ثبت درخواست جدید به گردش افتاده اند.
- ممکن است مالکان املاک پس از دریافت پروانه ساختمانی در مراحل مختلف (حتی پس از صدور پایان کار ساختمانی) اقدام به تخلف نمایند. لذا پس از هر ثبت درخواست جدید امکان مشاهده تخلف جدید وجود داشته و تعدد تخلفات مربوط به یک ملک محتمل می باشد.

۱-۴-۱-۲- انواع داده های موجود

داده های موجود از نوع کیفی (متنی و رتبه ای) و کمی (فاصله ای و نسبی) هستند. بعبارت دیگر در این مسئله با انواع داده ها سر و کار داریم.

بعنوان مثال :

فیلد «شرح نوع رای» دو وضعیت «بدوی» و «قطعی» را در خود جای داده است. (متنی)

فیلد «متراژسطح اشغال زمین» حاوی مقادیر مثبت عددی و پیوسته می باشد. (فاصله ای)

فیلد«تراکم» حاوی مقادیر مثبت عددی پیوسته که نمایانگر درصد می باشد.(نسبی)

فیلد «سهم مالکیت» حاوی مقادیر مثبت عددی پیوسته در بازه صفر تا شش می باشد.(نسبی)

فیلد« شماره ترتیب بازدید » حاوی اعداد صحیح یک تا ۸ که نمایانگر شماره بازدیدی است که اطلاعات از آن استخراج شده است.(رتبه ای)

۱-۴-۲- انتخاب و استخراج ویژگیها و پیش پردازش

پس از گردآوری داده ها، مطالعه بر روی تک تک ویژگیها صورت می پذیرد. هدف استخراج ویژگی این است که داده‌های خام به شکل قابل استفاده‌تری برای پردازش‌های آماری بعدی درآیند.در نگاه اول و با توجه به تعریف مساله بسیاری از فیلدها قابل حذف یا انتخاب هستند. مطالعه میدانی پایگاه

داده و گرفتن اطلاعات از کاربران نیز می تواند در انتخاب موثر باشد. مثلاً اینکه در بسیاری از مواقع مقادیر صفر در یک فیلد ثبت شده و تنها در شرایط خاصی مقدار غیر صفر دارد. انتخاب و استخراج ویژگی می تواند بر اساس اهمیت و میزان تاثیر، پراکندگی، آنتروپی صورت پذیرد. تعیین ویژگیهـای تاثیر گذار از موارد حائز اهمیت در ارزیابی نهایی خواهد بود.

در گامهای بعدی کارهای ذیل ممکن است در استخراج ویژگی انجام شود :

۱-۴-۲-۱- حذف نوفه یا نویز داده‌ها

۱-۴-۲-۲- جداسازی اجزای مستقل داده‌ها

۱-۴-۲-۳- کاهش ابعاد برای تولید بازنمایی مختصرتر

۱-۴-۲-۴- افزایش بعد برای تولید بازنمایی جدایی‌پذیری‌تر

استخراج ویژگی فرایندی متداول در انواع پردازش داده‌ها است. فرآیند های انتخاب و استخراج ویژگی و پیش پردازش، ارتباط زیادی با یکدیگر دارند. بطوریکه در برخی منابع گامهای یکی را در دیگری تعریف و تفسیر نموده اند. بعنوان مثال گامهای حذف نویز داده ها یا کاهش ابعاد را می توان نام برد.

مهمترین بخش از پیش پردازش داده ها شناخت ماهیت داده است. در این بخش با بررسی نوع داده، سیاستهای برخورد با داده های ناقص، پرت ، نویز و ناسازگار با توجه به خصوصیات هر ویژگی، تعیین شده، همچنین در صورت نیاز با بهره گیری از روشهای کاهش داده یا کاهش ویژگی های داده از جمله کاهش بعد، اقدامات مورد نیاز صورت خواهد پذیرفت و در نهایت و در صورت نیاز، با نرمال سازی -که بر اساس نوع داده ها متفاوت خواهد بود- ، آنها را برای مرحله عملیاتی آماده خواهیم نمود.

۱-۴-۳- خوشه بندی

خوشه بندی یکی از شاخه های یادگیری بدون نظارت می باشد و فرآیند خودکاری است که در طی آن، نمونه ها به دسته هایی که اعضای آن مشابه یکدیگر می باشند تقسیم می شوند که به این دسته ها خوشه گفته می‌شود. بنابراین خوشه مجموعه ای از اشیاء می باشد که در آن اشیاء با یکدیگر مشابه بوده و با اشیاء موجود در خوشه های دیگر غیر مشابه می باشند. برای مشابه بودن می توان معیارهای مختلفی را در نظر گرفت مثلا می توان معیار فاصله را برای خوشه بندی مورد استفاده قرار داد و اشیائی را که به یکدیگر نزدیکتر هستند را بعنوان یک خوشه در نظر گرفت که به این نوع خوشه بندی، خوشه بندی مبتنی بر فاصله نیز گفته می شود.

بعبارت دیگر خوشه بندی به عمل تقسیم جمعیت ناهمگن به تعدادی از زیر مجموعه ها یا خوشه - های همگن گفته می‌شود. وجه تمایز خوشه‌بندی از دسته‌بندی این است که خوشه‌بندی به دسته- های از پیش تعیین شده تکیه ندارد. در دسته‌بندی بر اساس یک مدل هر کدام از داده‌ها به دسته‌ای از پیش تعیین شده اختصاص می‌یابد؛ این دسته‌ها یا از ابتدا در طبیعت وجود داشته‌اند(مثل جنسیت، رنگ پوست و مثال‌هایی از این قبیل) یا از طریق یافته‌های پژوهش‌های پیشین تعیین گردیده‌اند.

در خوشه‌بندی هیچ دستهٔ از پیش تعیین شده‌ای وجود ندارد و داده‌ها صرفاً براساس تشابه گروه- بندی می‌شوند و عناوین هر گروه نیز معمولاً توسط کاربر تعیین می‌گردد. به طور مثال خوشه‌های علائم بیماری‌ها، ممکن است بیماری‌های مختلفی را نشان دهند و خوشه‌های ویژگی‌های مشتریان، ممکن است حاکی از بخش‌های مختلف بازار باشد. خوشه‌بندی معمولاً به عنوان پیش درآمدی برای بکارگیری تحلیل‌های داده‌کاوی یا مدل‌سازی به کار می‌رود.

در صورت بهره مندی از رده بندی داده هایی که با تعداد خوشه از قبل تعیین شده کار می کنند، یافتن تعداد مناسب خوشه اهمیت دارد. همچنین مطالعه معیارهای شباهت مختلف با توجه به ترکیبی بودن داده های مسئله حائز اهمیت می باشد.

۱-۴-۴- اعتبار سنجی خوشه ها

نتایج حاصل از اعمال الگوریتم خوشه‌بندی روی مجموعه داده ها با توجه به انتخاب پارامترهای الگوریتم، می‌تواند بسیار متفاوت از یکدیگر باشد. هدف از اعتبارسنجی خوشه‌ها یافتن خوشه‌هایی است که بهترین تناسب را با داده‌های مورد نظر داشته باشند. اینکار بر اساس معیارهای پایهٔ اندازه‌گیری برای ارزیابی و انتخاب خوشه‌های بهینه (تراکم درون خوشه ها و فاصله بین خوشه ها) صورت می پذیرد. انتخاب یک شاخص مناسب از بین شاخص های متعدد (از جمله شاخص دون ، شاخص دیویس بولدین و ...) مهمترین مرحله این بخش خواهد بود.

۱-۴-۵- تفسیر نتایج

خروجی کار به صورت گزاره های منطقی بیان می گردد. این گزاره ها با انضمام دانش مربوط به ضوابط و مقررات موجود و حتی فرآیند های نظارتی، مجموعه ای از پیشنهادات را برای شهرداری در پی خواهد داشت.

۱-۵- موانع و محدودیتهای پژوهش

۱-۵-۱- یکی از چالشهای پیش رو عدم ورود اطلاعات به صورت کامل برای هر رکورد می باشد ، که برای مرتفع نمودن آن، فرآیند تکمیل یا حذف داده به شرح فصل سه، شکل گرفته است.

۱-۵-۲- تنوع ورود اطلاعات مربوط به یک ملک حتی در یک دوره کوتاه مدت و بر اساس یک درخواست ثبت شده. بعنوان نمونه در برخی درخواستها اطلاعات مربوط به برخی فیلدها تکمیل نمی شود در حالی که ممکن است آن درخواست منتهی به ثبت تخلف شود. مثلاً تا زمانی که محاسبه عوارض برای یک ملک صورت نپذیرد قیمت منطقه ای برای آن ملک ثبت نخواهد شد. حال آنکه در برخی درخواستها اساساً نیازی به محاسبه عوارض وجود ندارد.

۱-۵-۳- کار با داده های ترکیبی یکی از محدودیتهای این مسئله هست. به عناون مثال عدد چهار در فیلدهای مختلف می تواند نمایانگر موراد ذیل باشد:

۱-۵-۳-۱- در فیلد تعداد طبقات می تواند نمایانگر چهارمین طبقه باشد. که داده ای رتبه ای بوده و وجود طبقات سوم و دوم را نیز در خود دارد.

۱-۵-۳-۲- در فیلد نوع خلاف معرف "بنای مسکونی، بدون پروانه، مازادبر تراکم، در کاربری مربوطه" می باشد.

۱-۵-۳-۳- در فیلد سهم مالکیت معرف نسبت مالکیت (تقریباً ۶۷ درصد) می باشد.

۱-۵-۳-۴- در فیلد مساحت عرصه، معرف متراژ عرصه ملک بعنوان کمیتی دو بعدی می باشد.

۱-۶- شرح واژه ها و اصطلاحات پژوهش

داده کاوی : داده کاوی ترجمه ی عبارت « Data Mining » و به معنای «کاویدن معادن داده» است.

تخلفات ساختمانی : تخلف ساختمانی عبارت است از بی اعتنایی به قانونمندیهای موجود در عرصه ساخت و ساز های شهری، قانون شکنی در ساختمان سازی و عدول از مقررات ساختمان سازی

داده های ترکیبی : ترجمه ی عبارت « Mixed Data » و به معنای داده هایی که همزمان شامل انواع عددی (فاصله ای و نسبی)، اسمی و طبقه بندی هستند.

فصل دوم

مبانی نظری و مرور روشهای مشابه

۲-۱- مقدمه

همانگونه که در فصل یک اشاره شد، مطالعات مربوط به تحلیل انگیزه های تخلفات ساختمانی در ایران عموماً بر پایه شیوه های مبتنی بر اخذ دانش از یک نمونه آماری مناسب و بر اساس مدلهای استاندارد رایج و تحلیل نتایج، صورت پذیرفته یا آنکه بر گزارشات آماری استخراج شده از بانک اطلاعات شهری استوار بوده است، و داده کاوی از بانک اطلاعات شهری که از طریق یادگیری بدون ناظر صورت پذیرد در ایران سابقه ای ندارد. اگر چه که پرداختن به داده کاوی (رده بندی) در داده های ترکیبی، موضوعی است که هر چند کم ، در حوزه های دیگر مورد توجه قرار گرفته است.

در این فصل و در بخش پیشینه پژوهش اینکه داده های ترکیبی چه داده هایی هستند و تفاوت آنها با داده های عددی چیست؟ یا رویکرد ها در مواجهه با این داده ها چگونه بوده است؟ مورد بررسی قرار خواهد گرفت. در ادامه برخی روشهای پیشنهادی مطرح شده و با مرور مبانی پیش پردازش داده ها، به شرح یک روش آسان و سریع خوشه بندی (الگوریتم لوید[8] موسوم به K-means) می پردازیم و سپس به الگوریتم های فراابتکاری خودکار[9] و کارکرد آنها (تنها برای ارزیابی عملکردمان) اشاره ای خواهیم نمود.

۲-۲- پیشینه پژوهش

مسعود یقینی و مهدی ورد در "خوشه بندی خودکار داده های مختلط با استفاده از الگوریتم ژنتیک" (8) به ارائه روشی پرداخته اند که با بهره گیری از یک معیار فاصله دقیقتر و الگوریتم ژنتیک به الگوریتمی با دقت بالاتر برسند.

[8]- Lloyd's algorithm
[9]- Automated meta-heuristic clustering

ساندرای دوی[۱۰] در مقاله " رده بندی داده ها با ویژگی های مختلط و بر اساس شباهت متریک واحد (9)" با بیان اینکه بسیاری از خوشه بندیها برای داده های صرفاً عددی یا داده های صرفاً طبقه بندی روشهای قابل اجرایی هستند و نه برای هر دو، به یک شکاف و فاصله غیر قابل قبول بین معیارهای شباهت متریک برای داده های طبقه بندی و داده های عددی اشاره نموده و خوشه بندی داده ها با صفات ترکیبی را با این روش، کار با ارزشی نمی داند. از نگاه ساندرای یک الگوریتم خوشه بندی کلی بر اساس شباهت اشیاء خوشه، باید طوری تنظیم شود که پاسخگوی داده های با ویژگیهای ترکیبی باشد.

شکی نیست که داده ها تنها زمانی معنی دار هستند که بتوانید اطلاعات مخفی را از دل آنها استخراج نمائید. با این حال، "مانع اصلی به دست آوردن دانش با کیفیت بالا از داده ها، ناشی ازمحدودیت های خود داده ها ست". این موانع عمده ی داده های جمع آوری شده ناشی از بزرگ شدن اندازه و حوزه های متنوع آن است.

برنامه های مختلف خوشه بندی در حوزه های گوناگون پدید آمده است. با این حال، بسیاری از الگوریتم های خوشه بندی مرسوم با تمرکز بر روی داده های عددی یا داده های طبقه بندی طراحی شده اند. در حالی که در دنیای واقعی داده ها اغلب حاوی هر دو نوع پارامتر عددی و طبقه بندی هستند. استفاده از این الگوریتم خوشه بندی سنتی به طور مستقیم برای این نوع از داده ها مشکل است. به طور معمول، زمانی که کاربران نیاز دارند الگوریتمهای خوشه بندی سنتی مبتنی بر فاصله را برای گروه بندی این انواع داده ها بکار بگیرند، یک مقدار عددی برای هر صفت داده طبقه بندی تعیین خواهند کرد.

برای مثال برخی از مقادیر طبقه بندی عبارتند از، "کم"، "متوسط" و "زیاد"، به راحتی می تواند به مقادیر عددی تبدیل شود. اما اگر صفات داده طبقه بندی شامل مقادیری شبیه : "قرمز"، "سفید" و

---

[۱۰]- M.Soundaryadevi1

"آبی"، و غیره باشد، نمی توان آن را به طور طبیعی مرتب نمود. در این شرایط، یک راه ساده برای غلبه بر این مشکل تبدیل مقادیر طبقه بندی به یکی از اعداد همان طبقه است، به عنوان مثال رشته باینری، و سپس اعمال روش های خوشه بندی مبتنی بر مقادیر عددی فوق الذکر. با این کار اطلاعات تشابه موجود در مقادیر طبقه نادیده گرفته شده و نمی تواند صادقانه ساختار شباهت از مجموعه داده ها را نشان دهد از این رو، برای حل این مشکل پیدا کردن یک متریک شباهت واحد و یکپارچه برای صفات طبقه بندی و عددی که در آن فاصله متریک بین داده های عددی و طبقه بندی را بتوان حذف نمود مطلوب است. پس از آن، یک الگوریتم خوشه بندی کلی که برای داده های عددی و طبقه بندی قابل اجرا است، که بر اساس این معیار واحد و یکپارچه ارائه شود.

در طول دهه گذشته، برخی کارها که سعی داشت تا یک متریک شباهت واحد و یکپارچه برای صفات طبقه بندی و عددی پیدا کند صورت پذیرفت. روشهای مختلف ارائه شده را می توان به دو گروه تقسیم نمود. در گروه اول الگوریتمها اساساً برای داده های کاملا طبقه بندی طراحی شده است، اگر چه آنها برای داده های ترکیبی و همچنین برای ویژگی های عددی که از طریق یک روش گسسته به یک طبقه بندی تبدیل گشته هم استفاده شده است. در ادامه برای این گروه، چندین روش بر اساس دیدگاه شباهت متریک، افراز گراف یا آنتروپی داده ها مطرح شده است.

لی و بیسوا[۱۱] "شباهت مبتنی بر خوشه بندی" (Similarity Based SBAC (10) Clustering)[۱۲] **Agglomerative** را ارائه نمودند. الگوریتمی که مبتنی است بر گودال شباهت متریک

---

[۱۱] - Li and Biswas

[۱۲] - در داده کاوی و آمار، خوشه بندی سلسله مراتبی (که اغلب hierarchical cluster analysis یا HCA نامیده می شود) یک روش آنالیز خوشه هست که به دنبال ساخت یک سلسله مراتب از خوشه ها است .استراتژی برای خوشه بندی سلسله مراتبی به طور کلی به دو نوع تقسیم می شوند:

جوش آتش فشانی(**Agglomerative**) : یک رویکرد "پائین به بالاست" که با در نظر گرفتن هر مشاهده بعنوان یک خوشه شروع می شود و جفت خوشه ها با هم ادغام شده و در ادامه به عنوان یک خوشه در نظر گرفته می شوند و به همین ترتیب تا حصول نتیجه پیش می رود.

تقسیمی (**Divisive**): یک رویکرد " بالابه پائین" است که با در نظر گرفتن همه مشاهدات بعنوان یک خوشه شروع می شود. و تقسیمات به صورت بازگشتی بسمت پائین سلسله مراتب انجام می شود.

که وزن بیشتری به مقادیر ویژگیهای غیر معمول اختصاص می دهد. مطابق محاسبه شباهت ها از توزیع آنها بدون دانش قبلی و در عین حال با تضمین ارزشهای ویژگی آنها.

SBAC الگوریتمی است که بر روی داده های ترکیبی با ویژگی های عددی و اسمی به خوبی کار می کند. یک الگوریتم Agglomerative یک درخت دیاگرام می سازد و یک اکتشاف واضح و ساده برای استخراج یک بخش از داده هاست. عملکرد SBAC بر روی داده های واقعی و مصنوعی تولید پایگاه داده ها بررسی شده است.

نتایج، اثربخشی این الگوریتم در کشف الگو بدون نظارت را نشان می دهد. مقایسه با سایر طرحهای خوشه بندی نشان دهنده عملکرد برتر این روش است. این روش دارای قابلیت های خوبِ برخورد با ویژگی های ترکیبی است، اما محاسبه آن کاملا دشوار است.

بیندیا ام ، خوزه تومی جی[۱۳] ، داده های دانشجویان را برای مشخص نمودن کارائی این روش خوشه بندی نموده اند.، در طول این سالها سوابق علمی هزار نفر از دانشجویان در موسسات آموزشی جمع شده و بسیاری از این داده های موجود در فرمت دیجیتال می باشد. با استخراج این حجم عظیم از داده ها ممکن است بینش عمیق تری به دست آورید. و می توانید برنامه ریزی رویکردهای آموزشی و استراتژی در آینده را مشخص تر سازید.

طرح آنها برای فرمول بندی این مسئله بود که بعنوان یک کار داده کاوی با استفاده از الگوریتم های خوشه بندی k-means و فازی C-means الگوی پنهان موجود در داده ها را بیابند و این دو الگوریتم را برای خوشه بندی داده های دانشجویان جهت تحلیل عملکرد بر اساس ورزیدگی شان مقایسه کنند.

[۱۳]- Bindiya M Varghese , Jose Tomy J

مینگ یی شیا[۱۴] ، "روش دو مرحله برای خوشه بندی ترکیبی داده های طبقه بندی و عددی" (11) ، در این روش آیتمها در ویژگیهای طبقه بندی پردازش شده، شباهتها یا روابط میان آنها بر اساس ایده هم–اتفاقی[۱۵] سازماندهی می شوند؛ پس از آن تمام ویژگی های طبقه بندی را می توان به ویژگی های عددی بر اساس این روابط ساخته شده تبدیل نمود. در نهایت، از آنجائیکه تمام داده های طبقه بندی به عددی تبدیل شده، الگوریتم های خوشه بندی موجود را می توان بدون زحمت برای پایگاه داده بکار برد. با این وجود، الگوریتم های خوشه بندی موجود از برخی از معایب و یا نقص روش دو مرحله ای رنج می برند.

روش دو مرحله ای ارائه شده، سلسله مراتب و اجزاء الگوریتم خوشه بندی را با اضافه کردن ویژگی ها به اشیاء خوشه کامل می کند. این روش روابط بین آیتم ها را تعریف کرده، و نقاط ضعف استفاده از الگوریتم خوشه بندی را بهبود می بخشد. الگوریتم TMCM الگوریتم های خوشه بندی HAC و-K means را برای خوشه بندی ترکیبی انواع داده ها کامل می کند. استفاده از سایر الگوریتم ها یا معیارهای شباهت پیچیده در TMCM ممکن است نتایج بهتری در پی داشته باشد .

۲-۳- اهمیت آماده سازی داده ها

اهمیت آماده سازی داده ها به دلیل این واقعیت است که؛ فقدان داده با کیفیت، برابر با فقدان کیفیت در نتایج کاوش است (12) و ورودی بد خروجی بد به دنبال دارد (13). وظیفه اصلی پیش پردازش داده ها سازمان دهی داده ها در شکل های استاندارد برای داده کاوی و یا سایر عملیات برنامه های کاربردی است. چه آنکه برای اعمال متدهای داده کاوی باید متناسب با آن داده ها را تهیه کرد.
مهمترین بخش از پیش پردازش داده ها شناخت ماهیت داده است. فلذا بحث را با انواع ویژگی در داده ها آغاز می کنیم.

---

[۱۴]- Ming-Yi ShihA

[۱۵]- Co-Occurrence

۲-۳-۱- انواع ویژگی ها در داده ها

۲-۳-۱-۱- عددی[۱۶]

۲-۳-۱-۱-۱- فاصله ای[۱۷]

خصوصیات :

- ✓ اندازه گیری در یک مقیاس از واحدهای هم اندازه
- ✓ مقادیر ترتیب دارند.
- ✓ عدم وجود نقطه صفر واقعی

مثال : تاریخهای تقویم ، درجه حرارت سانتیگراد و فارنهایت و در این مسئله فیلد ترتیب بازدید

۲-۳-۱-۱-۲- نسبی[۱۸]

خصوصیات :

- ✓ نقطه صفر ذاتی دارند.
- ✓ قابل بیان بصورت نسبی هستند.

مثال ۱ : درجه حرارت کلوین ، طول ، تعداد ، مقدار پول و در این مسئله فیلدهای «تراکم» و «تعداد طبقات زیرزمین» و «تعداد طبقات سایر» از این نوع هستند.

مثال ۲ : ۱۰ درجه کلوین دو برابر ۵ درجه کلوین است.

دسته بندی دیگری انواع ویژگی عددی را به دو دسته گسسته[۱۹] و پیوسته[۲۰] تقسیم می کند.

---

16 - Numeric
17 - Interval
18 - Ratio

۲-۳-۱-۲- **اسمی** : اسم دانشجو ، رنگ چشم ،گروه خونی، جنسیت، کدپستی، شماره دانشجویی در این مسئله تمامی فیلدهایی که با کلمه کد توضیح داده شده اند مثل «کد نوع اسکلت» یا «کد نوع سند». البته مقادیر مربوط به فیلد شماره بلوک نیز از این دست می باشند.

۲-۳-۱-۳- **طبقه بندی** : بعنوان مثال در این مسئله فیلدهای مربوط به «تعداد طبقات»

۲-۴- پیش پردازش[۲۱] داده ها

پس از گردآوری داده ها باید کیفیت آنها بررسی شود و با توجه به وابستگی نتایج عملیات داده کاوی به داده های مسئله، انتخاب نادرست و کیفیت پایین آنها ، صحت نتایج کار را به مخاطره می اندازد.

مهمترین عامل در یک پیش پردازش خوب ، فهم داده است. با کمک این موضوع، می توان مراحل بعدی عملیات داده کاوی را بهبود بخشید. به این معنی که می توان جامع و مانع بودن داده ها، هدف و کاربرد داده ها و مواردی از این دست را درک کرد تا ضمن افزایش قابلیت اطمینان به عملیات داده کاوی، سرعت انجام کار نیز افزایش یابد.

۲-۴-۱- عملیات اصلی پیش پردازشها داده ها

- پاکسازی داده ها[۲۲]
- یکپارچه سازی و تجمیع داده ها[۲۳]
- کاهش داده [۲۴]
- تبدیل داده ها [۲۵]

---

19 - Discrete Attribute
20 - Continuous Attribute
۲۱ - Preprocessing
۲۲ - Data Cleaning
۲۳ - Integration Data
۲٤ - Reduction Data

نکته : در پیش پردازش داده ها باید فرآیندها را بصورت ستون به ستون براساس مفهوم فیلد و مقادیر آن انجام داد.

۲-۴-۱-۱- پاکسازی داده ها

این مرحله شامل سیاستهای حل مشکل داده های بدون مقدار ، داده های آلوده به نویز و داده های ناسازگار می باشد.

۲-۴-۱-۱-۱- داده های بدون مقدار و مقادیری که **اندازه گیری** [۲۶] یا کامل[۲۷] **نشده اند.**

راه حل ها

۲-۴-۱-۱-۱-۱- حذف رکورد (یا فیلد) داده

البته در صورت نیاز به استفاده از این روش باید به یاد داشته باشیم که محاسبات انجام شده برای مقادیر آمار توصیفی؛ از قبیل میانگین، واریانس و کواریانس به اندازه های متفاوت نمونه مربوط خواهد شد که تأثیر آن باید مد نظر باشد.

برخی داده ها مفقود کاملاً از نظر آماری غیر وابسته به داده هایی است که تا کنون مشاهده شده اند ؛ این داده ها را مفقود شده ی کاملاً تصادفی[۲۸] می گویند. در برخی موارد نیز مقادیر مفقود، تصادفی هستند و به تعدادی از متغیرها یا طبقه داده های پیش بینی کننده مشروط می باشند. دسته ای دیگر از داده های مفقود نیز، غیر قابل چشم پوشی[۲۹] هستند؛ به این معنا که این نوع داده های مفقوده به کمک داده های مشاهده

---

[۲۵] - Data Transformation

[۲۶] - Missing Value

[۲۷] - Incomplete Value

[۲۸] - Missing Completely at Random (MCAR)

[۲۹] - No Ignorable Missing Data (NMD)

شده قبل از خود قابل نقل هستند .این قبیل تفاوت ها سبب می شود که روش های متفاوتی برای برخورد با مقادیر مفقود مورد استفاده قرار گیرد.

۲-۴-۱-۱-۱-۲- پر کردن با مقدار دلخواه (پر کردن به صورت دستی)

همان گونه که قابل پیش بینی هم می باشد این روش چندان عملی نیست؛ چرا که پیدا کردن و اصطلاحات لازم زمان بر است. البته در برخی مواقع تنها راه حل ممکن است. مثلاً، دو نام و آدرس فرضی محمد رحیمی ساکن تهران و محمدامین رحیمی ساکن تهران را در نظر بگیرید. اگر این دو نفر دقیقاً یکی بوده و تمامی سایر مشخصات آن ها نیز یکی باشند؛ تشخیص و رفع این مشکل ممکن است به کمک کامپیوتر مقدور نباشد. البته این موارد بسیار محدود است.

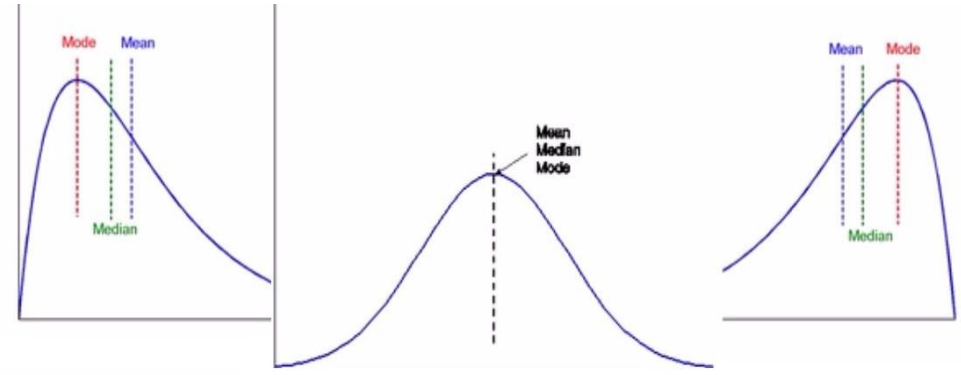
۲-۴-۱-۱-۱-۳- پر کردن به صورت خودکار

۲-۴-۱-۱-۱-۳-۱- پرکردن با مقدار ثابت سراسری

مسأله ای که در این صورت با آن مواجه خواهیم بود آن است که، ممکن است در حجم بالای داده ها ویژگی مقدار دهی شده با این مورد، جزء داده های محاسباتی محسوب شده و در محاسبات منظور گردد. و به این شکل ایجاد خطا نماید. به علاوه هنگامی که عملیات پاگ سازی داده ها برای ساخت انبار داده استفاده می شود، این روش انتخاب مناسبی نخواهد بود. در این مواقع با استفاده از یک فیلد به عنوان Flag برای مشخص کردن مقادیر ثبت نشده رکوردها در این فیلد (اگر ثبت شده بود فیلد Flag مقدار یک و در غیر اینصورت مقدار صفر داشته باشد)

۲-۴-۱-۱-۱-۳-۲- استفاده از یک معیار مرکزی (میانگین ، میانه)

- برای داده هایی که تقریباً بطور متقارن حول مرکز توزیع شده اند میانگین مناسب است. استفاده از این روش ممکن است سبب شود تا به دلیل تاثیر مقادیر نسبت داده شده به این ویژگی، نتایج به دست آمده به نفع این میانگین بایاس شود؛ حتی ممکن است اتخاذ این روش سبب حذف یا انتقال رکوردهای مربوط به یک دسته خاص از داده ها به سمت دسته نتایج دیگری شده و یک دسته مهم و واقعی از نتایج را نادیده بگیریم.
- اما اگر داده ها به یک سمت انحراف داشت و چولگی در آن بخش بیشتر بود آنگاه میانه شاخص مناسبتری خواهد بود.



شکل ۱,۲ : رابطه بین انتخاب روش پر کردن داده با معیار مرکزی و توزیع داده

فرمول تجربی :

Mean- Mode ≅ 3*(Mean-Median)

- ✓ در نمودار های طرفین که چولگی به یک سمت است میانه مناسبتر است (۵۰٪ داده ها در طرفین قراردارد)
- ✓ در نمودار وسط هر سه شاخص قابل قبول و قابل استفاده هستند.

نکته : محدودیت میانگین این است که الزاماً در بین داده ها موجود نیست ولی میانه حتماً در داده ها وجود دارد. به بیانی دیگر گسستگی داده ها در بسیاری از موارد موجب برتری انتخاب میانه نسبت به میانگین می شود.

مثال ۱: برای $x = [1\ 2\ 5\ 6\ 7]$ داریم $mean(x) = 4.200$ و $median =$ 5 مشاهده می کنید که میانگین در بین داده‌ها وجود ندارد.

مثال ۲ : در بسیاری از منابع از شاخص مرکزی زیر استفاده شده است :

$$\delta = \frac{f - f_{min}}{f_{max} - f_{min}}$$

۲-۴-۱-۱-۱-۳-۳- استفاده از یک معیار مرکزی مربوط به همان کلاس

مثال : پر کردن فشارخون یک فرد ۶۰ ساله با شاخص مرکزی کلاس همسالان او و نه شاخص مرکزی کل داده‌ها

۲-۴-۱-۱-۱-۳-۴- استفاده از مقادیر شایع یا پر احتمال[۳۰] (پر کردن با مقادیر با احتمال بیشتر)

این روش که پرکاربردترین روش قابل‌اعتماد است، شامل روش‌های استنتاجی و به‌کارگیری فرمول‌های بیزین، رگرسیون و درخت تصمیم است. به نوعی در این روش‌ها بر اساس استنتاج منطقی که مبتنی بر نوع اطلاعات موجود است؛ عمل پیش‌بینی صورت می‌گیرد.

- این مورد نیز مانند ردیف ج می‌تواند Class Based باشد.

---

۳۰ - Mode

نکته ۱ : مد برای انواع ویژگی‌ها قابل استفاده است.

نکته ۲ : تقریباً تمامی موارد ذکرشده در بالا تا حدودی مشکل بایاس[۳۱] یا تمایل به سمتی خاص را پیدا خواهند کرد.

باز هم بایستی یادآوری کنیم که ، نوع داده‌ها و شناخت آن‌ها قبل از پر کردن مقادیر مفقوده ضروری است. مثلاً **نمی‌توان داده طبقه‌بندی را با روش میانگین ویژگی پرکرد**، چرا که میانگین برای این نوع داده ها قطعاً بی معنا خواهد بود. درک این موارد در مواجهه با این قبیل مشکلات اهمیتی حیاتی دارد.

۲-۴-۱-۱-۲- **داده های آلوده به نویز**[۳۲] و دارای خطا

برخی خصوصیات نویز :

- ماهیت تصادفی دارد
- مقدار آن نست به داده اصلی از نظر قدر مطلق معمولاً خیلی کم و کوچک است
- تغییرات ناگهانی دارد

روشهای حذف نویز :

- رگرسیون (هموار سازی [۳۳])
- دسته بندی[۳۴]

---

[۳۱] - Bias

[۳۲] - Noisy Value

[۳۳] - Smoothing

[۳۴] - Bining

- (حذف «داده های پرت[۳۵]» [۳۶]) (این داده ها خصوصاً زمانی ایجاد می شوند که در زمانهای معدودی اندازه نویز زیاد باشد.)

۲-۴-۱-۱-۳- داده های ناسازگار[۳۷]

این مشکل مربوط به تناقض در داده ها بوده و از جمله مواردی است که نیازمند تجربه و صرف وقت بسیار است. به عنوان مثال وجود فرمت متفاوت در فیلد تاریخ تولد و سن مربوط به یک مشتری خاص، در صورتی که همخوانی لازم را نداشته باشد، ناسازگاری محسوب می شود. این گونه خطاها ممکن است به دلیل استفاده از منابع مختلف داده و در زمان ترکیب دو منبع مختلف داده ها روی دهد .

اما مسئلهٔ عمده ای که با آن مواجه می شویم و تشخیص آن بسیار مشکل است، تعیین ناسازگاری های نهفته است. به عنوان مثال اگر به دنبال کشف الگو در مورد مسائل مربوط به هتل داری باشید و قیمت مربوط به هتل های دنیا را از منابع مختلف جمع آوری کنید، جدای از بحث تبدیل نرخ ها و رفع ناسازگاری مربوط به مسایل خاص ارزی هر کشور، باز هم قیمت هتل ها نمی تواند ملاک مناسبی باشد، چرا که لازم است تا خدماتی همچون، صبحانه رایگان، استخر و سایر خدماتی را که در جاهای مختلف به شیوه های مختلف ارایه می شود، مد نظر داشت. به عبارتی قیمت هر شب اقامت در هتل در کنار نوع، شیوه و مقدار ارایه خدمات جانبی آن معنا پیدا می کند.

---

۳۵ - Outliers

۳۶ - disturbance Rejection

۳۷ - inconsistence

روش عمده و اصلی در حل ناسازگاری ها درک ماهیت داده ها است. اما در مواردی نیز ناسازگاری ها را که حاصل تجمیع چند منبع مختلف بوده و بیانگر افزونگی داده هاست، می توان با کمک روش های آماری بر طرف کرد.

۲-۴-۱-۲- یکپارچه سازی و تجمیع داده ها

دراین مرحله داده های ناهمگن پایگاه داده های مختلف را در یکجا مجتمع نموده و همگن می کنیم. مسایلی هم چون افزونگی داده ها در این دسته قرار می گیرند. اهم موارد مربوط به این بخش عبارتند از :

۲-۴-۱-۲-۱- اسامی متفاوت فیلدهای هم نوع

بعنوان مثال یک ویژگی با نام فیلدهای ID ، Shenase و SID در سه بانک مختلف با هم ادغام شود.

۲-۴-۱-۲-۲- حذف فیلدهای اضافه و وابسته

بعنوان مثال می توان با حذف ویژگی ضریب سلامتی ($\frac{w}{h^2}$) تنها از ویژگیهای قد (h) و وزن (w) بهره برد.

۲-۴-۱-۲-۳- داده تکراری

بعنوان مثال در تجمیع داده های مربوط به یک فروشگاه زنجیره ای قطعاً با مشتریانی مواجه می شویم که در بیش از یک بانک اطلاعاتی مشخصات دارند.

۲-۴-۱-۲-۴- ناهمخوانی مقداری

ممکن است در تجمیع بانک های اطلاعاتی مربوط به موسسات آموزشی مختلف با طبقه بندی متفاوت مواجه شویم بطوری که در یک بانک شاخص ارزیابی بین صفر تا بیست باشد

، در یکی بین صفر تا ۱۰۰ و در دیگری بر اساس چهار طبقه a ، b ، c و d ، که این مشکل با نرمال سازی[۳۸] مرتفع خواهد شد.

مثال : فرض کنیم مقادیر مربوط به بازه های $[۰,۴]$ ، $[-۱۰,۱۰]$ و $[۱۰,۲۰]$ را می خواهیم نرمال نمائیم. به ترتیب بازه اول را با تقسیم بر چهار ، بازه دوم را با جمع با عدد ۱۰ و آنگاه تقسیم بر ۲۰ و بازه سوم را با کم نمودن عدد ۱۰ و سپس تقسیم بر ده ، بر بازه $[۰,۱]$ نگاشت می دهیم.

۲-۴-۱-۳- کاهش داده ها

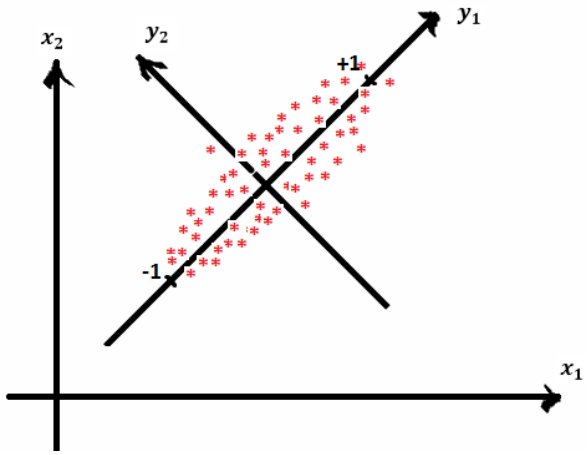
در این مرحله میتوانیم با روشهایی مثل خلاصه سازی یا استفاده از تکنیک های ارائه مختلف حجم داده را کاهش دهیم. این امر بخاطر محدودیتهای حافظه ، زمان و پردازش و جلوگیری از اتلاف منابع اهمیت دارد. نکته مهم در این مرحله از آماده سازی داده ها، آن است که دست یابی به نتایج تحلیلی مشابه با اصل و تمام داده ها تضمین گردد، چرا که در غیر این صورت این کاهش اثر مثبتی برای ما در پی نخواهد داشت. اهم موارد مربوط به این بخش عبارتند از :

۲-۴-۱-۳-۱- **کاهش ابعاد**

فرض کنیم در فضای (x_1 , x_2) داده هایی مشابه به شکل (۲,۲) داریم. با در نظر گرفتن مختصات جدید (y_1 , y_2) و صرفه نظر از تغییرات ناچیز در محور y_2 ، داده ها را از فضای دو بعدی (x_1 , x_2) به فضای $-1 \leq y_1 \leq 1$ نگاشت داده ایم. بدیهی است که بازیابی اطلاعات اولیه از داده های کاهش بعد یافته معمولاً به صورت کامل صورت نخواهد پذیرفت و در بسیاری از مواقع ممکن نخواهد بود.

---

[۳۸] - Normalization



شکل ۲,۲ : کاهش بعد ویژگی ها با در نظر گرفتن مختصات جدید و چشم پوشی از تغییرات ناچیز در راستای محور y_2

۲-۴-۱-۳-۲- انتخاب زیر مجموعه ای از ویژگیها[۳۹]

گاهی با انتخاب زیر مجموعه ای از ویژگیها و بهینه نمودن داده های نیازمند پردازش می توانیم داده ها را کاهش دهیم. انواع روشهای متداول عبارتند از :

a) Forward Selection F.W.S

فرض می کنیم فضای ویژگیها تهی است حال به ترتیب ویژگیهایی که بیشترین تاثیر را در تصمیم گیری دارند انتخاب می نمائیم.

b) Backward Selection B.W.S

فرض می کنیم فضای ویژگیها حاوی تمامی ویژگیهاست حال به ترتیب ویژگیهایی که کمترین تاثیر را در تصمیم گیری دارند حذف می نمائیم.

c) Hybrid Forward Backward Selection H.F.B.W

ترکیبی از دو روش فوق را بکار می گیریم.

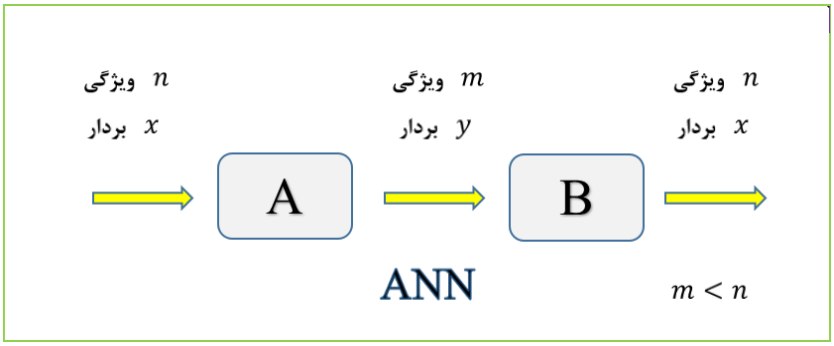
انواع الگوریتم های متداول کاهش بعد عبارتند از تحلیل مولفه اساسی (PCA)[۴۰]، NPCA، ICA ، منوفیلدها، شبکه های عصبی انجمنی (ANN)[۴۱] (شکل ۳,۲)

---

[۳۹] - Feature Subset Selection

[۴۰] - Principal components analysis

[۴۱] - Associative Neural networks



شکل ۳,۲ : ابعاد بردار ویژگیها در **ANN**

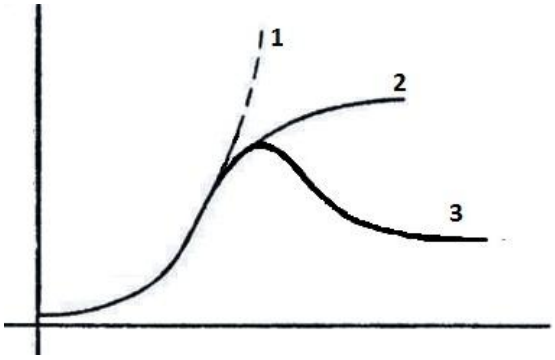
۲-۴-۱-۳-۳- کاهش تعداد

۲-۴-۱-۳-۳-۱- روشهای پارامتریک

مثال : مدلهای رگرسیونی

مثال : شرکتهای حاضر در بورس را می توان در یک دوره ۱۰ ساله به چند دسته تقسیم نمود و یک پارامتر را به عنوان نماینده هر دسته معرفی کرد. به عنوان نمونه می توان این شرکتها را به سه دسته تقسیم نمود:

- شرکتهایی که یک دوره شکوفایی را طی نموده و افول می کنند.(منحنی ۳ در شکل ۴,۲)
- شرکتهایی که به مرز اشباع شده گی می رسند.(منحنی ۲ در شکل ۴,۲)
- شرکتهایی که دائماً در حال رشد می باشند (منحنی ۱ در شکل ۴,۲)



شکل ۴,۲ : بدست آوردن الگوی طبقه ای که هر عضو به آن اختصاص دارد و نمایش آن عضو با آن الگو و پارامتر مختص خودش

۲-۴-۱-۳-۳-۲- روشهای غیر پارامتریک

از جمله :

- هیستوگرام : انتخاب یک یا چند نمونه (به نسبت وزن هر دسته) از هر دسته [۴۲] انتخاب می کنیم.
- خوشه بندی
- نمونه برداری
- شبکه های عصبی

۲-۴-۱-۳-۳-۳- فشرده سازی اطلاعات

در این خصوص دو روش با اتلاف[۴۳] و بدون اتلاف[۴۴] وجود دارد.

مثال ۱ : در تصویر تبدیلBMP به JPEG یا JPEG2000 و غیره

مثال ۲ : در صدا تبدیلWAV به MP3 و غیره

---

[۴۲] - Bine

[۴۳] - Lossy

[۴۴] - Loss Less

۲-۴-۱-۴- تبدیل داده ها

شاید بتوان تمام حالات قبل را را نوعی تبدیل نیز در نظر گرفت. در هر حال در برخی از موارد با عملیاتهایی چون هموارسازی و نرمال سازی میتوان داده های بی کیفیت را به داده های با کیفیت تبدیل کرد. اهم موارد مربوط به این بخش عبارتند از:

۲-۴-۱-۴-۱- هموار سازی

گاهی اوقات نیاز به هموار سازی نه بخاطر وجود نویز که بخاطر چشم پوشی از تغییرات کم اما واقعی داده ها صورت می پذیرد. که به آسان شدن پردازش کمک شایانی می نماید.

۲-۴-۱-۴-۲- استخراج [۴۵]یا ساخت ویژگی[۴۶]

بجای کار با همه داده ها می توان با بخش حائز اهمیت آنها کار کرد یا اینکه از تلفیق برخی خصوصیات داده می توان به ویژگی قابل قبولی دست یافت.

شکل ۵,۲ : استخراج ویژگی از داده ها بجای کار با کل داده ها

۲-۴-۱-۴-۳- تجمیع[۴۷]

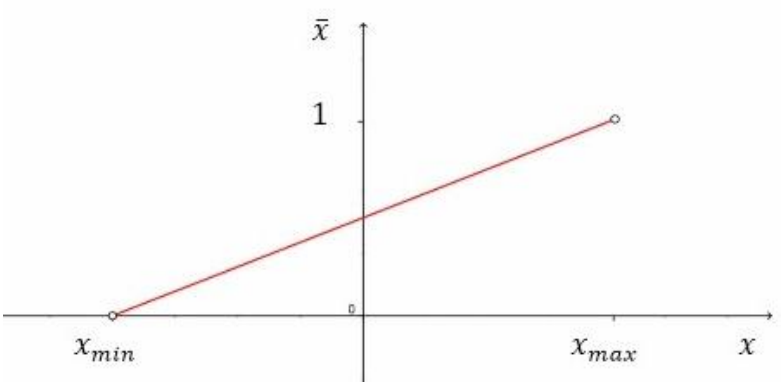
مثل تبدیل درآمد روزانه افراد به درآمد ماهانه جهت کاهش داده ها یا تبدیل بازه های ساعتی به بازه های روزانه

۲-۴-۱-۴-۴- نرمالسازی [۴۸]

۲-۴-۱-۴-۴-۱- خطی (شایع ترین روش)

[45] - Feature extraction
[46] - Feature Constraction
[47] - Aggregation
[48] - Normalization

فرض کنید $x_{min} \le x \le x_{max}$ می توان با تبدیل $\bar{x} = \frac{x - x_{min}}{x_{max} - x_{min}}$ آنرا به $0 \le \bar{x} \le 1$ نگاشت.



شکل ۶,۲ : نرمال سازی خطی و نگاشت هر بازه به بازه آماری $[۰ , ۱]$

۲-۴-۴-۱-۴-۲- آماری

مثال : نرمال سازی با میانگین صفر[۴۹]

می توان x را با تبدیل $\bar{x} = \frac{x - mean(x)}{std(x)}$ به $\bar{x}$ نگاشت به گونه ای که $mean(\bar{x}) = 0$ و $Var(\bar{x}) = 1$ شود.

۲-۴-۴-۱-۴-۳- مقیاس دِهیِ دَه - دَهی

فرض کنیم فیلد x بگونه ای باشد که به ازای هر x داشته باشیم $|x| \le 10^j$ (یعنی $-10^j < x < 10^j$) بطوریکه j کوچکترین عددی باشد که در این نامعادله صدق می کند. آنگاه با تبدیل $\bar{x} = \frac{x}{10^j}$ به بازه $0 \le \bar{x} \le 1$ نگاشته می شود.

۲-۴-۴-۱-۴-۴- نسبت فراوانی

در داده های اسمی برای هر حالت i با تعداد فراوانی f_i بر اساس فرمول زیر عمل نمائیم:

[49] - Zero Mean Normalization

$$\delta = \frac{f_i}{\sum f_i}$$

۲-۴-۱-۴-۵- گسسته سازی

می توان بجای نگاه پیوسته به کمیتها آنها را به صورت گسسته دسته بندی نمود. مثلاً زمان را به ۲۴ واحد یک ساعته در طول شبانه روز تقسیم نموده و داده های مربوط به هر بازه را درآن دسته بندی کرد.

۲-۴-۱-۴-۶- درختهای مفهومی

ممکن است برخی اطلاعات یک ویژگی حاوی مجموعه ای از اطلاعات نهفته باشد. به عنوان مثال زندگی در محله نارمک مبین زندگی در شهر تهران و کشور ایران و ... نیز هست. لذا می توان با در اختیار داشتن برخی ویژگیها به سلسله ای از خصوصیات رکورد دست یافت.

آنتروپی[۵۰]

آنتروپی درجه تصادفی بودن داده و مقدار اطلاع موجود در یک صفت را اندازه گیری می کند. و برد آن در بازه صفر تا یک می باشد.

$$Entropy = -\sum_{i \in outcomes} p_i \log_2 p_i$$

اگر همه اعضای مجموعه در کلاس یکسانی قرار گیرند Entropy=0 و اگر همه اعضا کاملاً تصادفی باشند Entropy=1 خواهد بود.

[۵۰] - Entropy

۲-۵- الگوریتم های خوشه بندی[51] در داده کاوی

جین، مورتی و فلین (۱۹۹۹) [52] ، در اثر تحقیقی خود با عنوان مروری بر خوشه بندی داده ها (14) ، بیان میکنند که خوشه بندی، طبقه بندی غیرنظارتی الگوها، شامل مشاهدات، اقلام دادها یا بُردارهای خصیصه ها، به گروهها و خوشه های منظم است.

نظم پنهان داده ها در سازمان برای تحلیل تجارت بسیار سودمند می‌باشد. خرده فروشی که می‌داند مشتریانش در چه گروه‌ی قرار دارند، می‌تواند فروش را بر اساس قاعدهٔ معینی هدایت کند.

الگوریتم های خوشه‌بندی (کلاسترینگ) اطلاعاتی را که ویژگی‌های نزدیک به هم و مشابه دارند را در دسته‌های جداگانه که به آن خوشه گفته می‌شود قرار می‌دهد. فرض کنید کودکی هستید که به همراه یک کیسه پر از تیله در اتاقی نشسته اید. اکنون کیسه را باز می‌کنید و اجازه می‌دهید تا تیله‌ها روی زمین حرکت کنند. متوجه می‌شوید که تیله ها رنگهای متفاوتی دارند. قرمز، آبی، زرد، سبز. تیله ها را برحسب رنگ جدا می‌کنید تا اینکه چهار گروه تیله داشته باشید. سپس متوجه می‌شوید که برخی از تیله ها بزرگ، بعضی کوچک و بعضی متوسط هستند. حال تصمیم می‌گیرید که تیله‌های بزرگ و کوچک را با هم و تیله‌های متوسط را به گروهی مجزا دسته بندی کنید. شما به این تقسیم بندی نگاه می‌کنید و از این کار راضی هستید. اکنون یک عملیات دسته بندی انجام داده‌اید.

دوباره نگاهی به خوشه‌ها می‌کنید و می‌بینید که نه تنها تیله هایی با رنگهای یکپارچه دارید، بلکه تیله‌های چشم گربه‌ای، شرابی، شیشه‌ای و احتمالا انواع دیگری نیز دارید. برخی از تیله‌ها دارای سائیدگی هستند. برخی از آنها دارای زوایایی هستند که بطور مستقیم حرکت نمی‌کنند. شما گروه بندی خود را براساس کدام خصوصیت انجام می‌دهید؟ اندازه، رنگ یا فاکتورهای دیگر از قبیل شکل

---

[51] - Clustering
[52] - Jain, Murty & Flynn

یا جنس؟ کدام یک بر دیگری ارجحیت دارد؟ اکنون سر درگم هستید آنقدر که به احتمال زیاد دوست دارید فقط بازی کنید!.

۲-۶- تعریف خوشه بندی

بخش بندی داده‌ها به گروه‌ها یا خوشه‌های معنادار به طوری که محتویات هر خوشه ویژگی‌های مشابه و در عین حال نسبت به اشیاء دیگر در سایر خوشه‌ها غیر مشابه باشند را خوشه‌بندی می‌گویند. خوشه بندی یک نوع طبقه بندی بدون سرپرست است یعنی اشیاء بر اساس شباهت‌هایی که باهم دارند تقسیم می‌شوند و نه بر اساس معیارهای از پیش تعیین شده. به گونه ای که می تواند تعداد دسته ها از قبل معین نباشد. دلیل دیگری آنکه به خوشه‌بندی، طبقه بندی بدون سرپرست می‌گویند. تحلیل این ویژگی ابزاری برای کشف ساختار داده ها است. در خوشه بندی با کشف شباهتها سعی می شود رفتارها شناسایی شده و امکان تصمیم گیری بهتر فراهم گردد.از این الگوریتم ها در مجموعه داده‌های بزرگ و در مواردی که تعداد ویژگی‌های داده زیاد باشد استفاده می‌شود.

بعبارت دیگر خوشه‌بندی ، اشیاء را براساس ویژگی‌هایی که با هم دارند گروه بندی می‌کند. هدف اصلی در خوشه بندی تقسیم بندی اشیاء به گونه‌ای است که در خوشه مخصوص خود دارای بیشترین شباهت و در مقایسه با اشیاء متعلق به خوشه‌های دیگر دارای بیشترین تفاوت هستند. با این معنا خوشه بندی را می توان یک مسئله بهینه سازی[۵۳] نیز قلمداد نمود.

[۵۳] - Optimization

۲-۶-۱- معیارمطلوبیت خوشه ها (معیارهای شباهت)

۲-۶-۱-۱- شباهت بالای هر دو شیء داخل یک خوشه

۲-۶-۱-۲- شباهت کم بین اشیاء هر دو خوشه متمایز

در خوشه‌بندی هر خوشه می‌تواند خود به چند زیر خوشه تبدیل شود. برای درک بهتر مشکلات تصمیم‌گیری برای تشکیل یک خوشه بندی به شکل ۷,۲ توجه کنید. در این شکل ۲۰ عضو که می‌توانند بصورت سه روش در خوشه بندی تقسیم شوند، نمایش داده شده است.

شکل ۷,۲ : ارزیابی سه خوشه بندی متفاوت از یک مجموعه ۲۰ عضوی

با توجه به شکل ممکن است که گرفتن چهار خوشه عقلانی نباشد (به علت شباهت نزدیک دو گروه)، به همین علت تاکید می‌کنیم که اشیاء خوشه را با توجه به وابستگی نوع داده‌ها و نتایج آن می‌توان بدست آورد. ایرادی که به برخی روشهای خوشه بندی ، که در آنها تعداد خوشه ها را از قبل مشخص می کنیم وارد است.

یکی از رایج ترین معیارهای شباهت فاصله می باشد. که بعنوان مثال فاصله هندسی اقلیدوسی[۵۴]، فاصله کسینوسی، correlation distance، Hamming distanc از معیارهای شباهت فاصله مورد استفاده در روش K-means نرم افزار مطلب می باشد.

محققین زیادی به بررسی محدودیتهای تحمیل شده از سوی این قسم معیارهای هزینه پرداخته اند. از معمولترین انتقادات وارده، بسیار ساده بودن این معیارها و عدم بازنمایی دقیق درک و تصور کاربر از ماهیت دادهٔ اصلی می باشد.

---

۵۴ - Euclidean

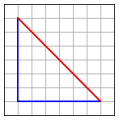
فاصله اقلیدوسی حالت خاصی از فاصله Minkowski است:

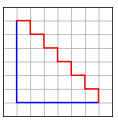
اگر $\mathbf{L_p}$ مبین فاصله مینکوسکی بین دو نقطه $P = (x_1 , x_2 \dots x_n)$ و $Q = (y_1 , y_2 \dots y_n) \in R$ باشد برابر است با :

$$L_P = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

در آنصورت :

- L_1 : فاصله cityblock یا Manhattan (P =1) و L_2 : فاصله اقلیدسی (P =2) است.


فاصله اقلیدسی


Cityblock ۱-۱-۱-۱-

- همچنین L_∞ : Chebyshev distance

$$\lim_{p\to\infty} \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} = \max_{i=1}^{n} |x_i - y_i|.$$ Type equation here.

$$\lim_{p\to\infty} (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}} = \quad |x_i - y_i|$$

۲-۷- ویژگیهای یک الگوریتم خوشه بندی مناسب

- مناسب بودن با انواع صفتها
- مقیاس پذیر
- ایجاد خوشه باشکل های مختلف
- حداقل بودن پارامترهای ورودی
- حساس نبودن به داده های نویز و پرت
- حساس نبودن به ترتیب داده های ورودی
- قابلیت تفسیرساده

همچنین برخی معایب که برای روش خوشه بندی شمرده اند به قرار زیر است :

- بررسی داده ها با ابعاد بالا نیازبه زمان بیشتری دارد.
- پاسخگوی همه نیازها نیست.
- معمولاً کارایی روش به فاصله تعریف شده وابسته است.
- قابلیت تفسیرهای مختلفی دارد.

۲-۸- انواع خوشه بندی

۲-۸-۱- **خوشه بندی اثباتی**: دراین روش اشیا را به دسته های مختلفی تقسیم بندی می کنیم (الگوریتم k-means)

۲-۸-۲- **روش سلسله مراتبی**: در این روش از دسته کوچک شروع کرده و خوشه بندی و دسته بندی انجام گرفته و سپس خوشه های نزدیک را با هم ترکیب نموده و خوشه های بزرگتر تشکیل می شود و همین کار تا ایجاد درخت خوشه ها ادامه می یابد. (*DBMS*) [۵۵]

۲-۸-۳- **خوشه بندی مبتنی بر چگالی** : در این روش نواحی که دارای چگالی زیاد می باشد به عنوان یک خوشه در نظر گرفته میشود.

۲-۸-۴- **روش مبتنی بر مدل** : در این روش فرض میکنیم داده ها از یک مدل پیروی میکنند و هدف از خوشه بندی ایجاد آن مدل میباشد.

۲-۸-۵- **مبتنی بر فراوانی** : در این روش برای پیدا کردن خوشه ها الگوریتمهای تکراری را بکار میبریم.

۲-۸-۶- **مشبک سازی** : در این روش فضاهای داده ای به مکعب هایی تقسیم شده و در هر کدام از آنها خوشه بندی را انجام می دهیم و در نهایت خوشه ها را بر اساس شباهت با یکدیگر ترکیب میکنیم.

۲-۸-۷- **روش افرازی** : دراین روش اشیا به k بخش افراز میشود هر افراز یک خوشه را تشکیل میدهد. خوشه ها با معیاری با نام تابع شباهت شکل می گیرد.

[۵۵] - Database Management System

۲-۹- الگوریتم k-means

در این الگوریتم k از ورودی گرفته شده و n شی موجود به k خوشه افراز می‌شود به طوریکه مربع خطاها در آن فرمول اقلیدسی و معیار شباهت کمتر یا حداقل باشد. دراین روش شباهت هر خوشه نسبت به متوسط اشیا آن خوشه سنجیده می‌شود و اصطلاحاً آن را مرکز خوشه می‌نامیم.

مراحل کار

- به طور تصادفی k نقطه دلخواه را به عنوان مراکز خوشه‌های ابتدایی انتخاب می‌کنیم.
- براساس مقدار شباهت هر نقطه (شیء) به مراکز خوشه‌ها نقطه‌ها را در خوشه ها قرار می‌دهیم. مرکز میانگین آن نقطه‌هاست.
- براساس خوشه‌های ایجاد شده مراکز جدیدی ایجاد می‌گردد.
- فرایند را از گام ۲ مرتباً تکرار می‌کنیم.
- فرآیند ۴ آنقدر تکرار می‌شود تا دیگر تغییری در خوشه‌ها ایجاد نشود.

محاسن روش خوشه‌بندی k-means را می توان پیاده سازی آسان و سرعت عملکرد دانست.

۲-۱۰- مشکلات روش خوشه‌بندی **k-means**

علی‌رغم اینکه خاتمه‌پذیری الگوریتم k-means تضمین شده است ولی الزاماً جواب نهایی آن واحد نبوده و همواره جوابی بهینه نمی‌باشد. به طور کلی این روش دارای مشکلات زیر است.

- جواب نهایی به انتخاب خوشه‌های اولیه وابستگی دارد.
- روالی مشخص برای محاسبهٔ اولیهٔ مراکز خوشه‌ها وجود ندارد.

- اگر در تکراری از الگوریتم تعداد داده‌های متعلق به خوشه‌ای صفر شد راهی برای تغییر و بهبود ادامهٔ روش وجود ندارد.
- در این روش فرض شده است که تعداد خوشه‌ها از ابتدا مشخص است. اما معمولا در عمل تعداد خوشه‌ها مشخص نیست.
- گرفتار شدن در دام بهینه محلی.

۲-۱۱- بررسی تکنیکهای اندازه‌گیری اعتبار خوشه‌ها

نتایج حاصل از اعمال الگوریتمهای خوشه‌بندی روی یک مجموعه داده با توجه به انتخاب‌های پارامترهای الگوریتمها می‌تواند بسیار متفاوت از یکدیگر باشد. هدف از اعتبارسنجی خوشه‌ها یافتن خوشه‌هایی است که بهترین تناسب را با داده‌های مورد نظر داشته باشند. دو معیارِ پایهٔ اندازه‌گیری پیشنهاد شده برای ارزیابی و انتخاب خوشه‌های بهینه عبارتند از :

- **تراکم**[۵۶] : داده‌های متعلق به یک خوشه بایستی تا حد ممکن به یکدیگر نزدیک باشند. معیار رایج برای تعیین میزان تراکم داده‌ها واریانس داده‌ها است. (از این ویژگی تحت عنوان پیوستگی[۵۷] یا فشردگی نیز نام برده اند)
- **تفکیک یا جدایی** [۵۸]: خوشه‌ها خود بایستی به اندازه کافی از یکدیگر جدا باشند. سه راه برای سنجش میزان جدایی خوشه‌ها مورد استفاده قرار می‌گیرد که عبارتند از:
  - فاصلهٔ بین نزدیک‌ترین داده‌ها از هر دو خوشه
  - فاصلهٔ بین دورترین داده‌ها از هر دو خوشه
  - فاصلهٔ بین مراکزخوشه‌ها

56 - Compactness
57 - Cohesion
58 - Separation

همچنین روش‌های ارزیابی خوشه‌های حاصل از خوشه‌بندی را به صورت سه دسته تقسیم می‌کنند که عبارتند از:

- معیارهای خروجی[۵۹]
- معیارهای درونی [۶۰]
- معیارهای نسبی [۶۱]

هم معیارهای خروجی و هم معیارهای درونی بر مبنای روش‌های آماری عمل می‌کنند و پیچیدگی محاسباتی بالایی را نیز دارا هستند. معیارهای خروجی عمل ارزیابی خوشه‌ها را با استفاده از بینش خاص کاربران انجام می‌دهند. معیارهای درونی، عمل ارزیابی خوشه‌ها را با استفاده از مقادیری که از خوشه‌ها و نمای آنها محاسبه می‌شود، انجام می‌دهند.

پایه معیارهای نسبی، مقایسهٔ بین شماهای خوشه‌بندی (الگوریتم به علاوهٔ پارامترهای آن) مختلف است. یک و یا چندین روش مختلف خوشه‌بندی چندین بار با پارامترهای مختلف روی یک مجموعهٔ داده اجرا می‌شوند و بهترین شمای خوشه‌بندی از بین تمام شماها انتخاب می‌شود. در این روش مبنای مقایسه، شاخص‌های اعتبارسنجی[۶۲] هستند. شاخص‌های ارزیابی بسیار متنوعی پیشنهاد شده‌اند که در این قسمت به شاخص DB و شاخص CS می پردازیم. برخی دیگر از شاخص های رایج اعتبار سنجی در مراجع (15) ، (16) ، (17) ، (18) ، (19) ، (20) ، (21) ، (22) معرفی شده و بعضاً مقایسه گردیده اند.

---

59 - External Criteria
60 - Internal Criteria
61 - Relative Criteria
62 - Validity-Index

۲-۱۲- شاخصهای اعتبارسنجی

شاخص‌های اعتبارسنجی برای سنجش میزان صحت[۶۳] نتایج خوشه‌بندی به منظور مقایسهٔ بین روشهای خوشه‌بندی مختلف یا مقایسهٔ نتایج حاصل از یک روش با پارامترهای مختلف مورد استفاده قرار می‌گیرند.

۲-۱۲-۱- شاخص دیویس بولدین[۶۴] (DB) – ۱۹۷۹

$$DB = \frac{\text{فاصله درون کلاسترها}}{\text{فاصله بین کلاسترها}}$$

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 که با فاصله دو مرکز بیان می شود (d_{ij}) تعریف می‌شود. شباهت بین دو خوشه را می‌توان به صورتهای مختلفی تعریف کرد و بایستی شرایط زیر را دارا باشد.

- $R_{i,j} \geq 0$
- $R_{i,j} = R_{j,i}$
- اگر s_i و s_j هر دو برابر صفر باشند آنگاه R_{ij} نیز برابر صفر باشد.
- اگر $S_j > S_k$ و $d_{i,j} = d_{i,k}$ آنگاه $R_{i,j} > R_{i,k}$
- اگر $S_j = S_k$ و $d_{i,j} < d_{i,k}$ آنگاه $R_{i,j} > R_{i,k}$

معمولا شباهت بین دو خوشه به صورت زیر تعریف می‌شود:

$$R_{i,q,t} = \max_{\substack{j \\ j \neq i}} \frac{S_{i,q} + S_{j,q}}{d_{i,j,t}}$$

(اگر دو خوشه تداخل داشته باشند بزرگتر از یک و اگر تداخل نداشته باشند کوچکتر از یک خواهد بود.)که در آن d_{ij} و s_i با روابط زیر محاسبه می‌شوند.

---

[63] - Goodness

[۶۴] - Davies Bouldin Index

فاصله دو مرکز دو کلاستر i و j : $d_{i,j,t} = \sqrt[t]{\sum_{p=1}^{d}|m_{ip} - m_{jp}|^t} = \|m_i - m_j\|_t$

پراکندگی اعضاء هر کلاستر : $S_{i,q} = \sqrt[q]{\frac{1}{N_i}\sum_{x \in C_i} d(x, m_i)^q}$

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

$$DB = \frac{1}{n_c}\sum_{i=1}^{n_c} R_{i,q,t}$$

که در آن $R_{i,q,t}$ به صورت زیر محاسبه می‌شود.

$$R_{i,q,t} = \max_{\substack{j \\ j \neq i}} \frac{S_{i,q} + S_{j,q}}{d_{i,j,t}} \qquad i, j = 1 \dots n_c$$

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

۲-۱۲-۲- شاخص چو ، سو و لای[۶۵](CS)

$$CS = \frac{\text{میانگین فاصله دورترین نقطه خوشه برای هر عنصر خوشه}}{\text{میانگین فاصله نزدیکترین خوشه به هر خوشه}}$$

فاصله دورترین نقطه از نقطه $x_p \in C_i$ برابر است با :

$$d_p^{max} = \max_{x_q \in C_i} d(x_p, x_q)$$

میانگین فاصله دورترین نقطه خوشه برای همه عناصر خوشه برابر است با :

$$\bar{d}_i = \frac{1}{N_i}\sum_{x_p \in C_i} d_p^{max} = \frac{1}{N_i}\max_{x_q} d(x_p, x_q)$$ (شاخصی از وسعت خوشه)

[۶۵] - Chou ,Su and Lai

و در سرانجام :

$$CS = \frac{\sum_i \frac{1}{N_i} \sum_{x_p \in C_i} \max_{x_q \in C_i} d(x_p , x_q)}{\frac{1}{k} \sum_i \min_{i \neq j} d(m_i , m_j)}$$

۲-۱۲-۳- معیار ارزیابی فیشر

معیاری با عنوان ضریب تغییرات که بر اساس آن هر چه اندازه این معیار کمتر باشد، می توان نتیجه گرفت که اعضای یک کلاس شباهت بیشتری با یکدیگر داشته و در عملیات خوشه بندی نتایج بهتری را حاصل نماید.با توجه به این معیار می توان الگوریتم خوشه بندی درست را انتخاب نمود.

$$f = \frac{m_v}{V_m} = \frac{\text{میانگین واریانسها}}{\text{واریانس مراکز}}$$

۲-۱۳- الگوریتم های فراابتکاری خودکار

در توابع هدف الگوریتم های خوشه بندی ساده معمولاً به دومین عامل مهم در خوشه بندی یعنی «شباهت کم بین اشیاء هر دو خوشه متمایز» توجهی نمی شود. از طرفی اگر چه خوشه بندی خود یک روش خودکار در تعیین اعضاء خوشه هاست ولی بطور معمول تعداد خوشه ها را باید از قبل تعیین نمود که این خود یک ایراد قابل توجه می باشد. سواگات داس و آجیت آبراهام[۶۶] در "خوشه بندی خودکار با استفاده از الگوریتم تکاملی اصلاح تفاضلی" (23) به معرفی یک روش خودکار برای

[۶۶] - Swagatam Das, Ajith Abraham

تعیین تعداد خوشه ها و با استفاده از شاخص های اعتبار[۶۷] پرداخته اند. که ضمن بررسی محل تشکیل خوشه ها، تعداد آنها را نیز کنترل می نماید. که شرح مختصری از روش آنها در ذیل می آید.

در این روش با کدینگ داده ها بصورت خطی و بر اساس جدول ۱,۲ :بصورت دو بخش مجزا در نظر گرفته شده است. بخش اول حاوی مقادیری با عنوان آستانه فعالسازی[۶۸] و بخش دوم حاوی مقادیر مراکز خوشه[۶۹] در این مرحله هستند. (کدینگ غیر خطی هم بشکل مشابه صورت می پذیرد)

جدول ۱,۲ : کدینگ خطی داده ها در روش داس و آبراهام

$\vec{V}_i(t) =$

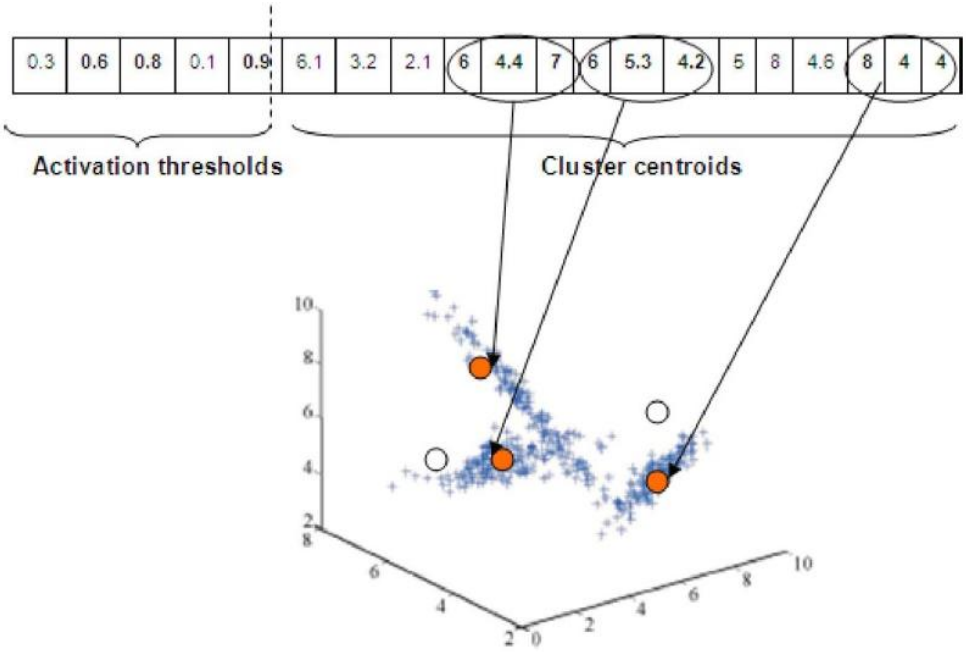
$T_{i,1}$	$T_{i,2}$	...	$T_{i,k_{max}}$	$\vec{m}_{i,1}$	$\vec{m}_{i,1}$	...	$\vec{m}_{i,k_{max}}$
Activation Threshhold				Cluster Centroids			

مقادیر آستانه بر اساس شاخصهای اعتبار سنجی DB و CS بدست می آیند. و روال کار آنستکه در صورتی که مقادیر آستانه از ۰,۵ بیشتر بود آن مرکز بعنوان یک مرکز فعال در نظر گرفته خواهد شد در غیر اینصورت چراغ مربوط به آن خاموش خواهد شد. در شکل ۸,۲ : چراغ مربوط به مقادیر ۰,۳ و ۰,۱ خاموش شده و مراکزی که شاخص آنها ۰,۶ ، ۰,۸ و ۰,۹ بوده به فعالیت خود ادامه می دهند.

[۶۷] - Validity Index

[۶۸] - Activation Thresholds

[۶۹] - Cluster Centroids



شکل ۸,۲ : خارج شدن مراکز دارای اعتبار کمتر از آستانه (مراکز سفید رنگ) از فرآیند خوشه بندی

فصل سوم

آماده سازی، پیش پردازش و رفع خطاهای دیتا

۳-۱- مقدمه

در مراجعه به یک بانک داده، قبل از هر چیز، هدف از مراجعه و خواستگاه مسئله مطرح است. فلذا در وحله اول انتخاب ویژگیهای اولیه، به مسئله طرح شده بر می گردد. میزان اثر گذاری ویژگیهای انتخابی در خروجی های مسئله، اثرات مخرب انتخاب ویژگیهای نادرست یا اضافه در رسیدن به جواب (از جمله ایجاد بایاس و انحراف از حل بهینه) و پاسخ به پرسشهای فرعی مطرح شده در فصل اول (چگونگی انتخاب ویژگیهای بیان کننده مسئله و مزیت ویژگیهای انتخاب شده بر سایر ویژگیها) موضوعاتی است که در فرآیند های پالایش و آماده سازی داده ها به آنها پرداخته می شود.

در این مسئله بر اساس مطالعه سازمان (شهرداری شاهرود) و قوانین موجود و اطلاع از برخی عوامل موثر در تخلفات (فارغ از سهم این ویژگی در وقوع تخلف)، در ابتدا به صورت شهودی نسبت به انتخاب ویژگیهای مسئله اقدام می نمائیم. اما می دانیم که داده های دنیای واقعی معمولاً ناقص، ناسازگار، دارای نویز، و مشکلات متعدد هستند. روالهای پالایش داده ها که به تفصیل در فصل دوم بدانها پرداخته شد سعی در پر کردن مقادیر مفقوده، حذف نویز، مشخص سازی اسکلت، اصلاح ناسازگاری و سایر موارد مورد نیاز برای آماده سازی داده ها دارند. در این فصل ابتدا به اهم موارد انتخاب یا عدم انتخاب ویژگیها اشاره خواهیم نمود. سپس روشهای پیاده سازی و تکنیک های به کار رفته پالایش داده ها را عنوان نموده، سیاستهای کار با داده های ترکیبی را بیان می نمائیم. و سرانجام با بیان شیوهٔ نرمال سازی داده ها فصل را به پایان خواهیم رساند.

۳-۲- اقدامات صورت پذیرفته جهت انتخاب ویژگیها

با عنایت به اینکه این پژوهش در پی یافتن عوامل موثر بر تخلفات ساختمانی است بررسی کلیه فیلدهای مربوط به املاک دارای خلاف در دستور کار قرار گرفته است. شاید در نگاه اول اینگونه به

نظر بیاید که با مراجعه به بانک، می توان کلیه اطلاعات یک ملک به همراه تخلفات آن را در یک رکورد تجمیع نمود در حالیکه اطلاعات مربوط به املاک ایستا نبوده و در طول زمان تغییر یافته اند و به عنوان مثال اگر تخلفات بصورت غیر همزمان صورت پذیرفته باشند، بسیاری از اطلاعات مربوط به آن ملک در تخلفات بعدی تغییر یافته اند فلذا و با عنایت به این موضوع موارد ذیل در چگونگی انتخاب ویژگیها مورد عمل قرار گرفته است:

۳-۲-۱- در فرآیند انتخاب ویژگی ها، با عنایت به اینکه تخلفات یک ملک بر اساس درخواستهای متفاوت، از زمان درخواست پروانه تا درخواست صدور پایان کار و یا حتی در طول سالهای مختلف بروز نموده است ، بهتر آن دیده شد که رکورد مربوط به هر درخواست با اطلاعات به روز شده آن ملک در همان درخواست به صورت یک رکورد جدا در نظر گرفته شود. فلذا امکان مشاهده چند رکورد برای پرونده مربوط به یک ملک وجود دارد.

۳-۲-۲- تنوع اطلاعات مربوط به یک ملک حتی در یک دوره کوتاه مدت و برای یک درخواست هم اجتناب ناپذیر می باشد. بعنوان مثال در برخی درخواستها اطلاعات مربوط به برخی فیلدها تکمیل نمی شود در حالی که ممکن است آن درخواست منتهی به ثبت تخلف شود. مثلاً تا زمانی که محاسبه عوارض برای یک ملک صورت نپذیرد قیمت منطقه ای برای آن ملک ثبت نخواهد شد[۷۰]. حال آنکه در برخی درخواستها اساساً نیازی به محاسبه عوارض نیست. این فیلدها حتی الامکان با قیمت منطقه ای مربوط به نزدیکترین زمان محاسبه عوارض به آن درخواست تکمیل گردیده است.

۳-۲-۳- از طرفی بدلایل متعدد از جمله خطای مامور بازدید یا اعتراض به رای، امکان دارد با یک درخواست دو سری اطلاعات مربوط به یک ملک (دو رکورد) درج شده باشد، لذا جهت اصلاح منحصر بفرد نبودن فیلد شماره درخواست، بجهت دسترسی های بعدی به مشخصات ملک و جلوگیری از درهم ریختگی داده ها در فرآیند پیش پردازش (انواع مرتب سازیهایی که صورت می پذیرد)، فیلدی با عنوان کلید به داده ها افزوده شده که منحصر به فرد بوده و در هر زمان امکان مرتب سازی داده ها به شکل اول را میسر می سازد.

---

[۷۰] – قیمت منطقه ای (ارزش معاملاتی زمین-موضوع ماده ۶۴ قانون مالیاتهای مستقیم) که در محاسبه مبلغ جریمه بکار گرفته نمی شود. بلکه ملاک ارزش معاملاتی ساختمان موضوع تبصره ۱۱ ماده ۱۰۰ قانون شهرداریهاست.

۳-۲-۴- یکی از ایده های مطرح در آغاز کار جهت مشخص نمودن نوع تخلف مربوط به هر رکورد بعد از انتخاب ویژگیهای ضروری و پیش پردازش اولیه، افزودن برداری با بعد ۲۳ (به تعداد نوع تخلف) به انتهای بردار ویژگی بود بگونه ای که مقادیر تک تک ویژگیهای آن صفر یا یک باشد. بعبارت دیگر در صورت وقوع هر تخلف در آن رکورد مقدار مربوط به آن تخلف یک و در غیر اینصورت صفر در نظر گرفته شود. اما با عنایت به بند «ت» و عدم امکان تجمیع تخلفات یک ملک در یک رکورد، تصمیم بر آن گرفته شد تا تحت عنوان فیلدی واحد، تخلفات کد شود. بعبارت دیگر هر رکورد مربوط به یک درخواست از یک ملک و تنها حاوی یک نوع تخلف خواهد بود.

۳-۲-۵- چالش دیگر پیش رو عدم امکان تفکیک موقعیت ملک، تنها بوسیله فیلد قیمت منطقه ای می باشد. زیرا بنا به روال موجود شهرداری شاهرود، قیمت منطقه ای بصورت دستی و توسط کاربر از دفترچه تعرفه عوارض انتخاب می شود. لذا بدلیل مشابهت در عدد مربوط به این عامل در مناطق مختلف امکان تفکیک دقیق موقعیت املاک میسر نخواهد بود. بعبارت دیگر املاک دارای قیمت منطقه ای یکسان در جای جای شهر پراکنده بوده و مکان یابی املاک تنها براین اساس میسر نیست. لیکن با در نظر گرفتن فیلد شماره بلوک، موقعیت منطقه ای ملک، بصورت حدودی قابل تفکیک می باشد.

۳-۲-۶- کد نمودن مقادیر کیفی در دستور کار قرار گرفته تا بردار ویژگی مربوط به هر رکورد تنها حاوی اعداد باشد.

۳-۲-۷- ترکیبی بودن داده های مسئله از مسائل حائز اهمیت می باشد. همانگونه که در فصل دو بدان پرداخته شد دو رویکرد در مقابله با این داده ها وجود دارد.

- رویکرد اول با پذیرش کامل نبودن کدینگ دیتا صرفاً بر اساس تخصیص اعداد ترتیبی، با پیش پردازش (و بعضاً با بکار گیری معیارهای شباهت و شاخص های اعتبارسنجی اصلاح شده) نقش این اعداد و تفاوت هایشان در تعیین خوشه ها را بهبود می بخشند.

- در رویکرد دوم اهمیت کدینگ داده ها حائز اهمیت بوده و خوشه بندی را تنها با بهبود معیارهای شباهت و شاخص های اعتبارسنجی، کار موثری نمی داند. این رویکرد بدنبال اصلاح روشهای کدینگ داده یا پردازشهای چند مرحله ای است.

در این پژوهش خروجی هر دو رویکرد (با اعمال کدینگ مربوطه برای هر کدام) مقایسه خواهد گردید.

۳-۳- بررسی ویژگی های موجود

با عنایت به مکانیزه شدن شهرداری شاهرود از سال ۱۳۸۷ ، داده کاوی بر روی اطلاعات بانک مربوط به سالهای ۱۳۸۷ تا ۱۳۹۴ این شهر مشتمل بر حدود هفده هزار رکورد تخلفات ساختمانی (ارجاعی به کمیسیون ماده ۱۰۰) حاوی فیلدهای در دسترس ذیل الذکر صورت می پذیرد.
(توضیح اینکه در کلیه جداول، سلول مقادیر ویژگی که در هیچ رکوردی از بانک ثبت نشده به صورت رنگی درآمده است)

- فیلد نوع رای حاوی مقادیر «بدوی» و «قطعی» و مقادیر نامشخص است.
- شماره ترتیب بازدید – یکی از تفکیک کننده های اطلاعات رکورد های مربوط به یک ملک می باشد. (این مقدار در بانک بین یک تا هشت ثبت گردیده است)

- نوع سند – شامل یازده گونه اطلاعات

جدول ۱,۳ : نوع سند

۱. ثبتی منگوله دار	۲. ثبتی ماده ۱۴۷	۳. وکالتنامه	۴. اجاره نامه
۵. وقف نامه	۶. رهنی	۷. هبه نامه	۸. قرارداد واگذاری
۹. قولنامه ای	۱۰. صلح قطعی	۱۱. نامشخص	

- بخش ثبتی: در شاهرود دو بخش ثبتی یک و دو وجود دارد. که با رکوردهای نامشخص سه وضعیت خواهد شد.
- مساحت طبق سند (مقداری عددی بین صفر تا ۳۸۶۲۱۴,۵ متر)
- نوع زمین شامل شش گونه اطلاعات :

جدول ۲,۳ : نوع زمین

۱. بایر	۳. دایر مشجر	۵. بایر محصور
۲. موات	۴. دایر ساختمانی	۶. مزروعی

- مرحله ساختمانی

جدول ۳,۳ : مرحله ساختمانی

۱. فونداسیون	۴. سقف همکف	۷. سقف طبقه دوم	۱۰. سقف طبقه سایر	۱۳. اتمام کار
۲. اسکلت بندی	۵. سقف نیم طبقه	۸. سقف طبقه سوم	۱۱. سفت کاری	۱۴. نامشخص
۳. سقف زیر زمین	۶. سقف طبقه اول	۹. سقف طبقه چهار	۱۲. نازک کاری	

- نوع اسکلت

جدول ۴,۳ : نوع اسکلت

۱. بتونی	۳. تیر چوبی	۵. نیمه فلزی	۷. نیمه بتونی	۹. نامشخص
۲. فلزی	۴. آجری	۶. خشت و گل	۸. صنعتی ساز	

- وضعیت ساختمان

جدول ۵,۳ : وضعیت ساختمان

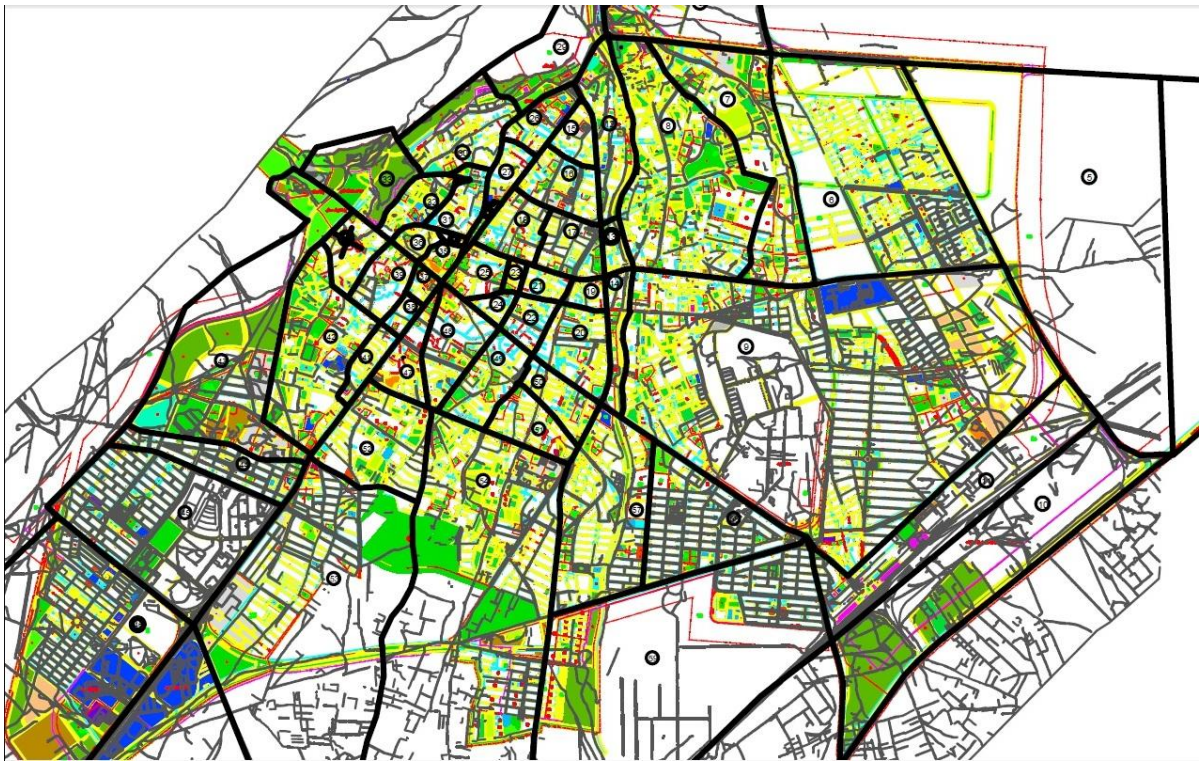
۱. مسکونی	۱۴. سینما	۲۷. صندوق قرض الحسنه اقتصادی	۴۰. آشپزخانه
۲. مسکونی تجاری	۱۵. بانک	۲۸. ندامتگاه	۴۱. کتابخانه
۳. تجاری	۱۶. گرمابه	۲۹. زمین مزروعی	۴۲. تجاری پرزباله
۴. فرهنگی	۱۷. فروشگاه	۳۰. باغ	۴۳. آموزشی دولتی
۵. آموزشی	۱۸. پارکینگ	۳۱. پارک	۴۴. بیمارستان
۶. ورزشی - تفریحی	۱۹. انبار	۳۲. پست برق	۴۵. درمانگاه
۷. گردشگری - پذیرائی	۲۰. کارگاه	۳۳. مسکونی آپارتمانی	۴۶. نامشخص
۸. پزشکی درمانی	۲۱. کارخانه	۳۴. مخروبه مسکونی	
۹. حمل و نقل	۲۲. بلا استفاده	۳۵. کاروانسرا	
۱۰. اداری دولتی	۲۳. تجاری مسکونی	۳۶. زمین بایر	
۱۱. انجمن ، شورا ، کانون	۲۴. صندوق و موسسه مالی	۳۷. مدرسه	
۱۲. مذهبی	۲۵. صندوق قرض الحسنه	۳۸. شهرداری	
۱۳. تاریخی	۲۶. اوقاف	۳۹. پروژه	

- تعداد طبقات زیرزمین : عددی صحیح (بین صفر تا سه) + یک حالت نامشخص
- تعداد طبقات سایر : عددی صحیح (بین صفر تا هشت) + یک حالت نامشخص
- شماره بلوک : عددی صحیح (بین یک تا شصت و یک) + یک حالت نامشخص
- شماره ردیف (شرح و موقعیت محل) : مقداری به صورت متنی که عمدتاً تلفیق دو عدد با فرمت های متفاوت که مبین محل دقیق ملک و عرض شارع مجاور می باشد.
- بر جبهه (ضریب محله) : مقداری عددی صحیح (یک و دو)
- تراکم : عددی صحیح (بین صفر تا ۱۸۰)
- تراکم تجدید نظر : مقداری عدد صحیح (بین دوازده تا ۱۱۸۰) + مقادیر متنی "بافت ویژه" ، "بافت قدیم" ، "بافت" ، "با نظر میراث" + یک حالت نا مشخص
- حد نصاب تفکیک : عددی صحیح (بین صفر تا ۷۰۰۰) + یک حالت نا مشخص
- محدوده

جدول ۶,۳ : محدوده

۱. خدماتی	۲. قانونی	۳. استحفاظی	۴. بافت ویژه	۵. نامشخص

- درصد اشغال زمین : عددی صحیح (بین صفر تا ۱۰۰) + یک حالت نا مشخص
- سهم مالکیت : عدد بین یک تا شش
- قیمت منطقه ای : عدد صحیح (بین ۱۴۵ تا ۱۶۰۰۰۰) + یک حالت نا مشخص



شکل ۱,۳ : (تصویری از نقشه شاهرود با تفکیک منطقه ای مناطق مختلف)

- نوع خلاف

جدول ۷,۳ : نوع خلاف

نوع خلاف	فراوانی
۱. بنای مسکونی، دارای پروانه، درحد تراکم، در کاربری مربوطه	۱۵۴۰
۲. بنای مسکونی، دارای پروانه، مازادبر تراکم، در کاربری مربوطه	۵۸۴۶
۳. بنای مسکونی، بدون پروانه، درحد تراکم، در کاربری مربوطه	۱۳۰۸
۴. بنای مسکونی، بدون پروانه، مازادبر تراکم، در کاربری مربوطه	۱۰۲۰
۵. بنای مسکونی، دارای پروانه، درحد تراکم، در کاربری مغایر	۶
۶. بنای مسکونی، دارای پروانه، مازادبر تراکم، در کاربری مغایر	۶۰
۷. بنای مسکونی، بدون پروانه، درحد تراکم، در کاربری مغایر	۱۱
۸. بنای مسکونی، بدون پروانه، مازادبر تراکم، در کاربری مغایر	۷۱۰
۹. کسری پارکینگ	۱۲
۱۰. بنای غیرمسکونی، دارای پروانه، درحد تراکم، در کاربری مربوطه	۳۸۴۵
۱۱. بنای غیرمسکونی، دارای پروانه، مازادبر تراکم، در کاربری	۱۱۴
۱۲. بنای غیرمسکونی، بدون پروانه، درحد تراکم، در کاربری مربوطه	۱۹۸
۱۳. بنای غیرمسکونی، بدون پروانه، مازادبر تراکم، در کاربری مربوطه	۲۵۶
۱۴. بنای غیرمسکونی، دارای پروانه، درحد تراکم، در کاربری مغایر	۱۱۶
۱۵. بنای غیرمسکونی، دارای پروانه، مازادبر تراکم، در کاربری مغایر	۷
۱۶. بنای غیرمسکونی، بدون پروانه، درحد تراکم، در کاربری مغایر	۸۳۷
۱۷. بنای غیرمسکونی، بدون پروانه، مازادبر تراکم، در کاربری مغایر	۶
۱۸. بنای بدون مجوز مبتنی بر کاربری مصوب در حریم شهر	۴۴۱
۱۹. بنای بدون مجوز مغایر بر کاربری مصوب در حریم شهر	۳
۲۰. تجاوز به حریم	۳۳۱
۲۱. سایر	۶۳
جمع	۱۶۷۳۰

- نوع رای

جدول ۸,۳ : نوع رای

نوع رای	فراوانی
عوارض	۲۱۷
جریمه	۱۲۵۴۵
تعطیل	۵
اعاده	۵۲۰
تامین پارکینگ	۴۷
اصلاحی پخی	۰
تعویق رای	۰
تخریب	۴۶۳
رفع تعرض	۳
برائت	۱۴۴۴
نامشخص	۱۴۸۶
جمع	۱۶۷۳۰

۳-۴- پیش پردازش داده ها

عملیات پیش پردازش در محیط اکسل و با استفاده از توابع آن صورت پذیرفته است. عدم انجام این بخش در برنامه های اجرائی، یا در نرم افزار های رایج داده کاوی و صرف وقت جهت تبدیل داده ها بدلیل آن بوده که با کلیه داده ها بر اساس یک روال ثابت برخورد نگردیده و بعنوان مثال برای نرمال سازی، داده ها به سه بخش تقسیم شده اند. (داده هایی که نیازی به نرمال سازی ندارند ، داده های اسمی ، داده های عددی و داده های نسبی) از طرفی شیوه کدینگ داده های اسمی و نسبی (نوع دوم)، در نرم افزاری های موجود ممکن نبود.

۳-۵- سیاست های کار با داده های ترکیبی

همانگونه که در فصول قبل اشاره شد، افزایش بُعد برای تولید بازنمایی و تفکیک هر چه بهتر ویژگیها از اقدامات مربوط به استخراج ویژگی است. از طرفی و بر اساس موارد اشاره شده در پیشینه تحقیق در فصل دوم، موضوعی که مورد انتقاد برخی از محققین واقع گردیده ، کدینگ داده های اسمی به صورت تخصیص اعداد صحیح ترتیبی است. چه آنکه بعضاً هیچ ترتیبی در دنیای واقعی و بلحاظ مفهومی بین مقادیر ممکن و قابل تخصیص در داده های اسمی وجود ندارد.

۳-۶- شیوه دیگر کدینگ داده های اسمی

می توان ویژگیهای اسمی ، طبقه بندی و حتی ویژگیهای نسبی[۷۱] را با این روش اینگونه بیان نمود که هر ویژگی اسمی را به تعداد مقادیر کاندید برای آن ویژگی در بانک، توسعه داده و بدین ترتیب بجای یک ویژگی، به تعداد مقادیر ممکن برای آن، بردار ویژگی را توسعه می دهیم. حال در یک رکورد تنها به ازای فیلد مربوط به مقدار واقعی عدد یک و به ازای سایر مقادیر صفر را در نظر می گیریم. بعنوان مثال در ذیل مقادیر سه رنگ قرمز ، آبی و زرد در نظر گرفته شده اند. با این روش حتی رنگهای بینابینی را می توان بصورت تلفیقی ارائه نمود.

جدول ۹,۳ : کدینگ داده های اسمی

کدینگ داده های اسمی		
فیلد زرد	فیلد آبی	فیلد قرمز
۰	۰	۱
۰	۱	۰
۱	۰	۰

کدینگ ساده	رنگ
۱	قرمز
۲	آبی
۳	زرد

[۷۱] - در نظر داشته باشید که ویژگیهای عددی نسبی هم رفتار متفاوتی از ویژگیهای عددی دارند بعنوان مثال ویژگی عددی مقدار ۶۰ روی نمودار یک نقطه است در حالی که ویژگی عدد نسبی ۶۰ شامل سطح تمامی مقادیر صفر تا ۶۰ می باشد.

۳-۶-۱- توسعه شیوه عنوان شده برای کدینگ داده های نسبی

یک پیشنهاد برای مقادیر نسبی نیز آنست که می توان مثلاً برای فیلد سهم مالکیت بجای کد کردن

عدد ۴ در ازای مالکیت چهاردانگ، مقادیر فیلدهای یک ، دو ، سه و چهار را یک در نظر گرفت. این کار کاملاً متفاوت از کدینگ معمولی و بدلیل افزایش بعد مسئله، بدون تاثیر در رفتار ویژگیها در ابعاد قبلی خواهد بود. می توان برای مقادیر ۰,۵ هم تنها یک فیلد مجزا در نظر گرفت تا از افزوده شدن چندین فیلد جلوگیری نمود.

جدول ۱۰,۳ : کدینگ پیشنهادی برای داده های نسبی

سهم مالکیت	کدینگ ساده	کدینگ پیشنهادی						
		۱	۲	۳	۴	۵	۶	۰,۵
۱	۱	۱	۰	۰	۰	۰	۰	۰
۲	۲	۱	۱	۰	۰	۰	۰	۰
۳	۳	۱	۱	۱	۰	۰	۰	۰
۴	۴	۱	۱	۱	۱	۰	۰	۰
۵	۵	۱	۱	۱	۱	۱	۰	۰
۶	۶	۱	۱	۱	۱	۱	۱	۰
۱,۵	۷	۱	۰	۰	۰	۰	۰	۱
۵,۵	۸	۱	۱	۱	۱	۱	۰	۱

۳-۷- اهم موارد مورد استفاده در پیش پردازش داده های مسئله

۳-۷-۱- مرحله پاکسازی داده

۳-۷-۱-۱- بدلیل ناکافی بودن داده ها (عدم ورود اطلاعات) در اکثر رکوردها فیلدهای ذیل حذف گردیدند.

- درصد اشغال زمین
- متراز سطح اشغال زمین
- تعداد طبقات مجاز
- تعداد واحد اضافه
- نوع نما

۳-۷-۱-۲- بدلیل آنتروپی پائین ویژگی محدوده نیز حذف می گردد. زیرا عملاً تاثیری بر خوشه بندی نخواهد داشت.(عموم املاک ثبت شده در محدوده قانونی هستند) همچنین انتظار داریم که سهم مالکیت که عموماً بصورت شش دانگ ثبت شده است، نقش چندانی در خوشه بندی ایفا نکند.

۳-۷-۱-۳- همچنین بدلیل تنوع محتوا و نیز ترکیب موارد و عدم انتخاب کاربران از یک لیست ثابت از فیلد شرح کاربری صرف نظر گردید. ذکر این نکته ضروریست که مشاور طرح تفصیلی، عموماً وضعیت ساختمان را در تدوین طرح لحاظ نموده است که بعنوان فیلدی در داده ها موجود است.

۳-۷-۱-۴- با عنایت به اینکه فیلدهای قیمت منطقه ای و شماره بلوک در داده ها موجود است و به صورت ترکیبی مبین موقعیت ملک هستند از کار با فیلد شماره ردیف (شرح و موقعیت محل) نیز صرف نظر گردید.

۳-۷-۱-۵- در کلیه فیلدها یکسان سازی فرمت ورودی داده ها صورت پذیرفته است. مثلاً در فیلد «تراکم تجدید نظر» مقادیر «*180 - 120*» ، «*180 -120*» «*120180*» ، «*120 - 180*» ابتدا همگی با فرمت یکسان درج شده اند.

۳-۷-۱-۶- در فیلدهای حاوی داده های اسمی مقدار صفر بعنوان نا مشخص در نظر گرفته شده است. (پر کردن با مقدار ثابت سراسری)

۳-۷-۱-۷- برای مواردی (از قبیل مثال ردیف ۲) نیز از میانگین استفاده شده است. یعنی بجای «*180 - 120*» عدد ۱۵۰ در نظر گرفته شده است.

۳-۷-۱-۸- در موارد بسیار محدودی که فیلد اسمی بوده نیز از معیار مد در کلاس داده برای مقادیر نامشخص استفاده شده است.

۳-۷-۱-۹- داده های ناسازگار

در فیلدهایی چون فیلد «تراکم تجدید نظر» با انواع داده های «NULL » ، « بافت ویژه»، « بافت قدیم» ، «بافت» ، «میراث» و «با نظر میراث» و ... مواجه بودیم که بدلیل نسبت بسیار کم این داده ها (۳۴۱ رکورد) و همچنین عدم دسترسی به مقدار عددی محتمل یا قطعی برای آنها وضعیت همه نامشخص(برابر صفر) در نظر گرفته شد.

همچنین رکوردهای حاوی اعدادی چون ۹۰ برای سایر طبقات یا ۲۴ برای طبقات زیر زمین که غیر واقعی بوده و خطای کاربر می باشد نیز بطور کامل حذف گردیده است.

۳-۷-۲- نرمال سازی داده ها

داده های اسمی که به روش دوم کد گردیده اند نیازی به نرمال سازی نداشته اند. سایر داده های مسئله به دو صورت نرمال گردیده اند.

- هر داده عددی x بر مبنای فرمول زیر نرمال شده است :

$$\boldsymbol{\delta} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- هر داده نسبی، به نسبت سهمش از کل (مثلاً نسبت سهم در شش دانگ مالکیت) در بازه $[\mathbf{0}, \mathbf{1}]$ نگاشته می شوند.

- برای تنها یک حالت داده اسمی (فیلد شماره بلوک) برای هر حالت $\boldsymbol{i}$ با تعداد فراوانی $\boldsymbol{f_i}$ بر اساس فرمول زیر نرمال شده اند:

$$\boldsymbol{\delta} = \frac{\boldsymbol{f_i}}{\sum \boldsymbol{f_i}}$$

۳-۸- اهم نتایج اقدامات آماده سازی– پیش پردازش و رفع خطاهای دیتا

در نهایت پس از اعمال رویه های پیش پردازش فوق الاشاره، ۱۶۷۳۰ رکورد حاصل گردید که با رویکرد اول هر رکورد حاوی ۲۱ ویژگی (کدینگ نوع یک) و در رویکرد دوم هر رکورد حاوی ۱۱۲ ویژگی (کدینگ نوع دو) می باشد.

پیش پردازش در دو وضعیت صورت پذیرفته است :

۳-۸-۱- ضریب منطقه ای شامل مقادیر صفر نیز باشد. حالتی که اساس پژوهش را تشکیل می دهد و از بسیاری از رکوردها صرفنظر نمی نماید. در این حالت یکی از فیلدهای موثر در موقعیت ملک (ضریب منطقه ای) با روشهایی از قبیل پر نمودن با نزدیکترین مقدار، مقدار دهی شده و همچنین تعداد کمی از رکوردهایی که امکان پیش بینی مقدار برای آنها وجود نداشت به صورت نامشخص در فرآیند خوشه بندی شرکت داده شده اند. لازم به ذکر است که ویژگیِ همیشه دارای مقدارِ شماره بلوک که حاوی اطلاعات موقعیت حدودی ملک می باشد هم در فرآیند خوشه بندی دخالت دارد.

۳-۸-۲- ضریب منطقه ای مخالف صفر باشد.

همانگونه که در فصل سه نیز بدان اشاره شد تنوع اطلاعات مربوط به یک ملک حتی در یک دوره کوتاه مدت و برای یک درخواست هم اجتناب ناپذیر می باشد. بعنوان مثال در برخی درخواستها اطلاعات مربوط به برخی فیلدها تکمیل نمی شود در حالی که ممکن است آن درخواست منتهی به ثبت تخلف شود. مثلاً تا زمانی که محاسبه عوارض برای یک ملک صورت نپذیرد قیمت منطقه ای برای آن ملک ثبت نخواهد شد[۷۲]. حال آنکه در برخی درخواستها اساساً

[۷۲] – قیمت منطقه ای (ارزش معاملاتی زمین-موضوع ماده ۶۴ قانون مالیاتهای مستقیم) که در محاسبه مبلغ جریمه بکار گرفته نمی شود. بلکه ملاک ارزش معاملاتی ساختمان موضوع تبصره ۱۱ ماده ۱۰۰ قانون شهرداریهاست. (بخاطر داشته باشید که دلیل استفاده از ویژگی قیمت منطقه ای در کنار ویژگی شماره بلوک، تعیین حدود و موقعیت املاک بوده است)

نیازی به محاسبه عوارض نیست. یک شیوه برخورد با این رکوردها، حذف آنهاست. اگـر چـه کـه سیاست در این پژوهش حفظ حداکثری اطلاعات بانک بوده، برخی نتایج حاصل از حذف حدود شش هزار رکورد با مقادیر صفر برای قیمت منطقه ای به شرح ذیل می باشد.

۳-۸-۲-۱- این رکوردها همه زمین از نوع دایر ساختمانی هستند.

۳-۸-۲-۲- نوع سند این رکوردها فقط شامل ثبتی منگوله دار ، ثبتی ماده ۱۴۷ ، اجاره نامه یا صلح قطعی است.

۳-۸-۲-۳- سهم مالکیت رکوردهای دارای خلاف تماماً شش دانگ می باشد.

۳-۸-۲-۴- همه املاک در محدوده قانونی هستند.

۳-۸-۲-۵- تمامی آراء صادره از نوع جریمه هستند.

۳-۸-۲-۶- تخلفات از انواع ردیفهای دو ، سه ، چهار ، پنج ، شش، هفت ، هشت و ده هستند.

۳-۸-۲-۷- بلوکهای یک تا پنج و پنجاه و یک تا شصت و یک فاقد رکورد هستند. بعبارت دیگر تخلف ثبت شده ندارند.

۳-۸-۲-۸- همه املاک فاقد زیر زمین هستند.

۳-۸-۲-۹- نوع اسکلت از ردیف یک تا چهار می باشد.

۳-۸-۲-۱۰- از نظر مرحله ساختمانی در مرحله سقف طبقه چهار ، سقف طبقه سایر ، سفت کاری ، نازک کاری یا اتمام کار قرار دارند.که عدم وجود سقفهای تا طبقه سوم معنا دار به نظر می رسد.

فلذا در صورت استفاده از این داده ها در عملیات خوشه بندی فیلدهای تک مقدار (یا فیلدهایی که غالباً دارای یک مقدار می باشند قابل حذف به نظر می رسند).

فصل چهارم

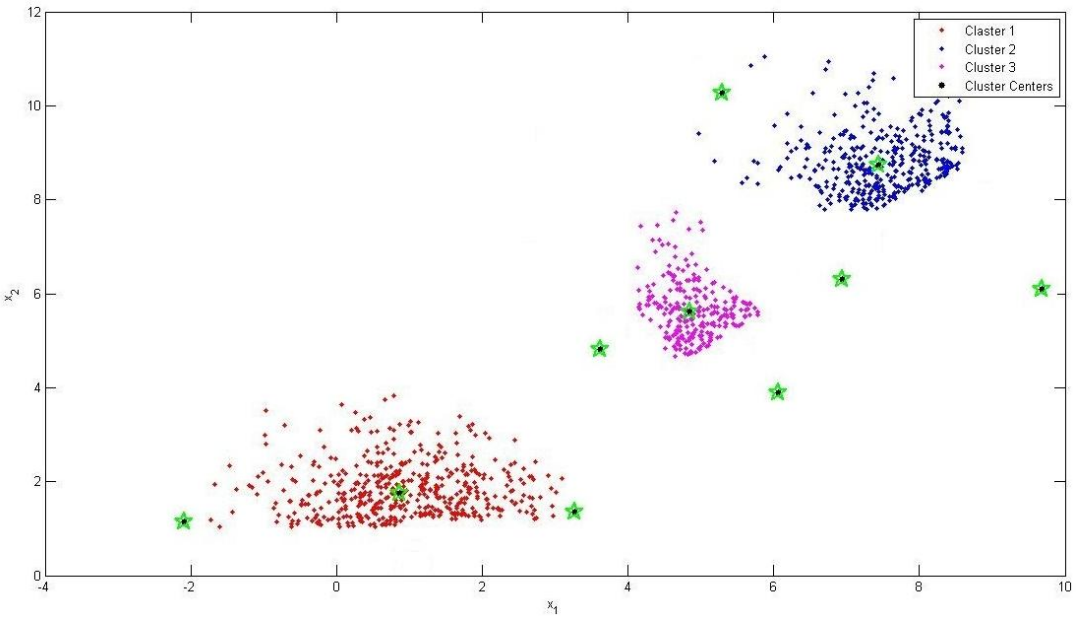
روشهای پیشنهادی – پیاده سازی و نقد

۴-۱- مقدمه

در فصل دوم ضمن آنکه به مبانی نظری پژوهش و کارهای مشابه پرداختیم به تفصیل فرآیندهای شناخت ماهیت داده ها و گردآوری آنها ، سیاست های مربوط به آماده سازی و پیش پردازش و خوشه بندی اشاره نموده و در فصل سوم نسبت به اعمال روشهای آماده سازی و پیش پردازش اقدام نمودیم. در این فصل ابتدا به روش های تعیین تعداد خوشه ها و اعتبار خوشه ها خواهیم پرداخت و سپس ضمن اعمال روش K-means بر داده ها، نتایج را با سایر روشها مقایسه می نمائیم. توضیح اینکه کار را با دو نوع کدینگ نوع یک (بر اساس تخصیص مقادیر عددی به داده های اسمی) و کدینگ نوع دو (بر اساس روش دوم مطرح شده در فصل سوم) انجام خواهیم داد. تحلیل تفاوتها پایان بخش این فصل خواهد بود.

۴-۲- آیا **K-Means** خود می تواند تعیین کننده تعداد خوشه ها باشد؟

در برخورد با یک نمونه خاص (شکل ۱,۴) این فرض را مطرح نموده ایم که در صورت بزرگتر بودن K از تعداد واقعی خوشه ها (به عنوان مثال M)، تعداد K-M خوشه را تنها از میان داده های پرت و با تعداد اندک اعضاء انتخاب خواهد کرد. در این صورت می توانیم بر اساس سنجش تعداد اعضاء تخصیص یافته به هر خوشه و تعیین یک آستانه کوچک برای حذف خوشه هایی که تعداد اعضایی کمتر از مقدار آستانه، به خود اختصاص داده اند، نسبت به انتخاب K مناسب اقدام نمائیم.



شکل ۱,۴ : تعداد واقعی خوشه ها سه تا است و K=10 در نظر گرفته شده است، تعداد هفت خوشه تک عضو یا با اعضاء محدود ایجاد شده اند.

بررسی این فرض را با یک آزمایشی شهودی انجام می دهیم. الگوریتم K_Means را برای ۳۰۰۰ داده دو بعدی تصادفی ایجاد شده توسط تابع randn (تابع توزیع نرمال با میانگین صفر و انحراف معیار یک)پیرامون سه نقطه X1=(1,1) ، X2=(2,3.5)، X3=(2,2) (برای در نظر گرفتن مقداری تداخل در داده ها) با مقادیر K=4 و K=10 و K=6 در ذیل آمده است:

جدول ۱,۴ : اجرای K_Means با K=10 برای ۳۰۰۰ داده دو بعدی تصادفی حول سه نقطه

خوشه	1	2	3	4	5	6	7	8	9	10	جمع
فراوانی	220	339	315	305	237	358	271	203	357	395	3000

جدول ۲,۴ : اجرای K_Means با K=6 برای ۳۰۰۰ داده دو بعدی تصادفی حول سه نقطه

خوشه	1	2	3	4	5	6	جمع
فراواز	421	543	615	526	379	516	3000

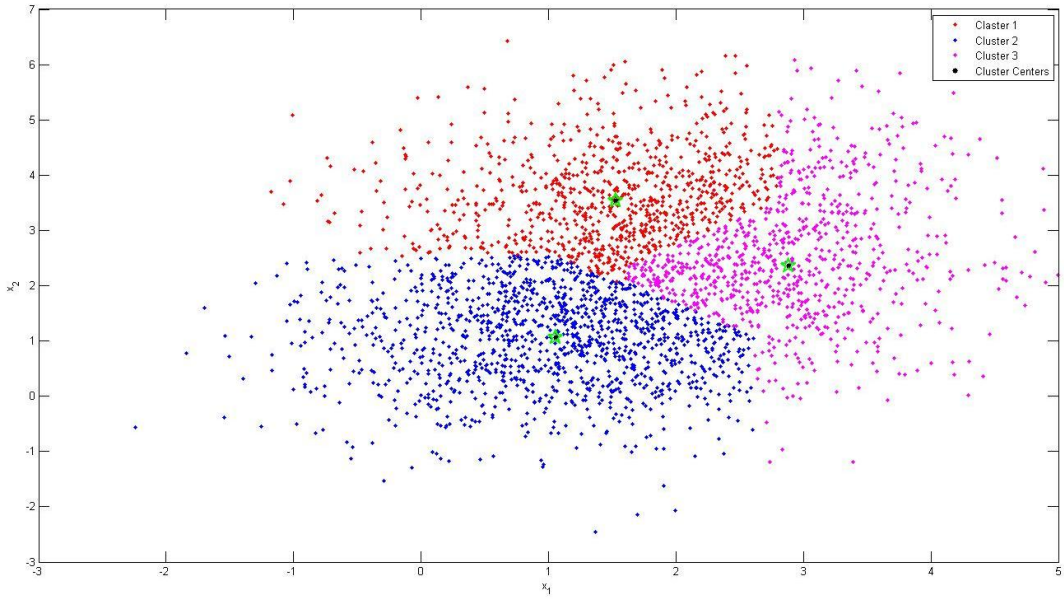
جدول ۳,۴ : اجرای K_Means با K=6 برای ۳۰۰۰ داده دو بعدی تصادفی حول سه نقطه

خوشه	1	2	3	4	جمع
فراواز	624	689	792	895	3000

توزیع درست و حل مسئله (شکل ۲,۴ :) :

جدول ۴,۴ : اجرای K_Means با K=3 برای ۳۰۰۰ داده دو بعدی تصادفی حول سه نقطه

خوشه	1	2	3	جمع
فراواز	879	1279	842	3000



شکل ۲,۴ : توزیع درست خوشه ها در مثال شهودی

توزیع متناسب داده ها در بین خوشه ها در هر سه آزمایش (K=4 ، K=6 و K=10) بیان گر نا معتبر بودن این فرض و غیر قابل تکیه بودن آن می باشد. توضیح اینکه در تکرار روش، برای داده های آزمایش، نتایج مشابهی بدست آمده است. همچنین از مقایسه جواب مسئله با حالتهای سه گانه قبل به این نتیجه می توان رسید که در حقیقت خوشه های واقعی به دو یا چند خوشه توسعه یافته اند که علی رغم وجود شباهت حداکثری در داخل خوشه ها، دارای تفاوت حداکثری در بین خوشه ها نیستند. و آن نمونه بخصوص که فرض ما را شکل داده بود،

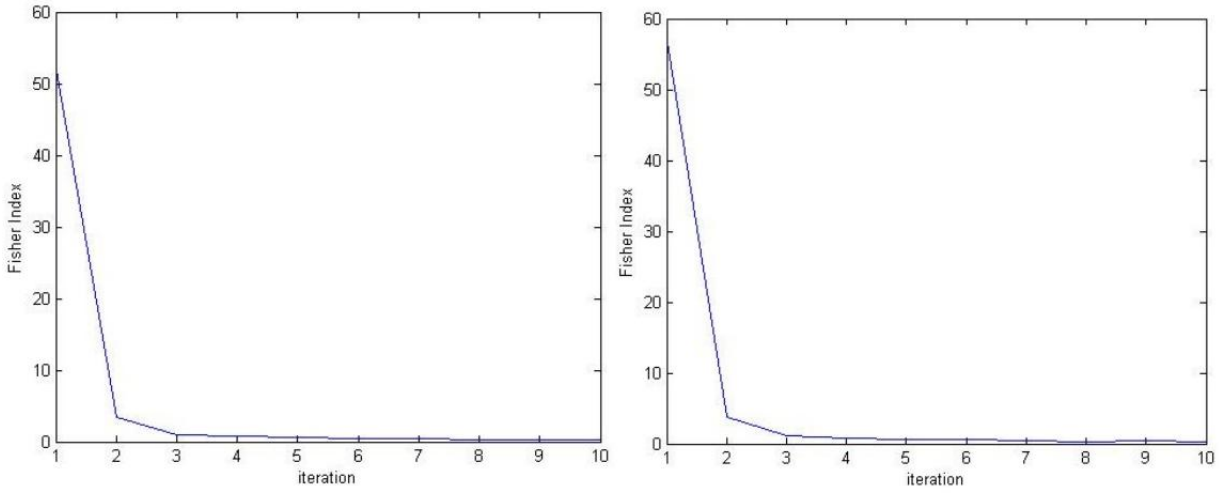
محصول یکی از مشکلات روش K-Means بوده است.[۷۳] با عنایت به سایر موارد تحقیقی در این پژوهش اتفاقاً این موضوع را می توان ضعف بزرگ الگوریتم لوید دانست.

۴-۳- بررسی معیار ارزیابی فیشر

به خاطر دارید که معیار فیشر با رابطه زیر بیان می شد.

$$f = \frac{m_v}{V_m} = \frac{\text{میانگین واریانسها}}{\text{واریانس مراکز}}$$

نتیجه چند بار اجرای K-Means و تعیین معیار ارزیابی فیشر در هر تکرار بر همان داده های آزمایش قبلی در شکل ۳,۴ :نمایش داده شده است.

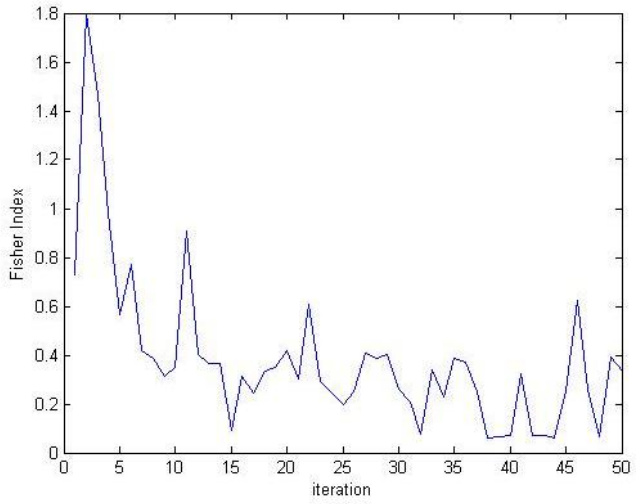


شکل ۳,۴ : نتیجه دو بار اجرای معیار فیشر روی داده های ایجاد شده به روش آزمایش قبل

با توجه به اینکه ساختار داده ها توسط خود ما برای سه نقطه X1=(1,1)، X2=(2,3.5)، X3=(2,2) با توزیع نرمال ایجاد شده بود، با توجه به خروجی شکل ۴,۳ که از تکرار ۳ به بعد مقادیر ناچیزی را به خود اختصاص داده است، و ما را به انتخاب سه خوشه هدایت می کند، به نظر می رسد که این معیار

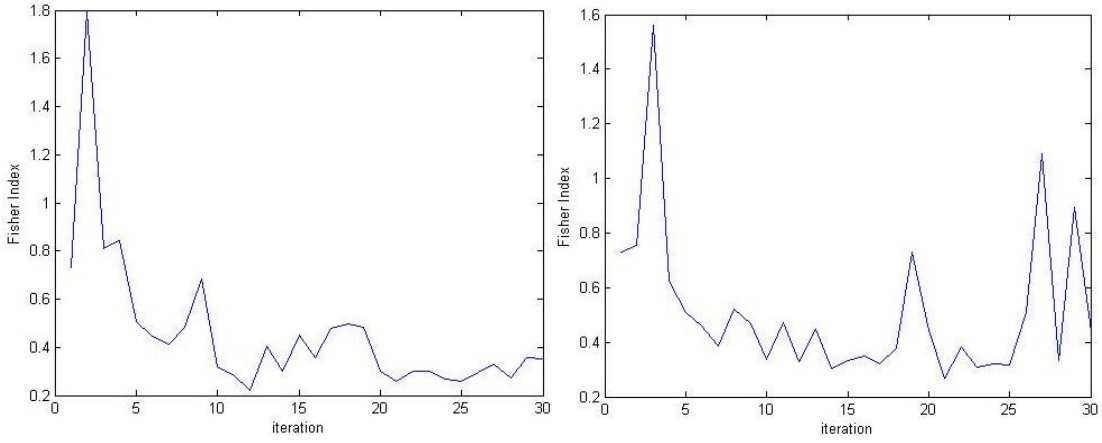
[۷۳] - اگر در تکراری از الگوریتم تعداد داده‌های متعلق به خوشه‌ای صفر شد راهی برای تغییر و بهبود ادامهٔ روش وجود ندارد.

ارزیابی، برای مسئله ما قابل اتکا باشد. لذا در ادامه خروجی برنامه را بر اساس این معیار بر روی هر دو نوع کدینگ بررسی می کنیم.

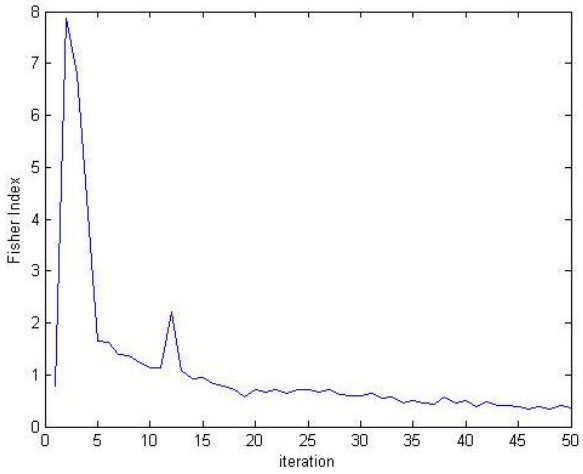


شکل ۴,۴ : ارزیابی K-Means با معیار فیشر برای K=1...50 برای داده های با کدینگ نوع یک

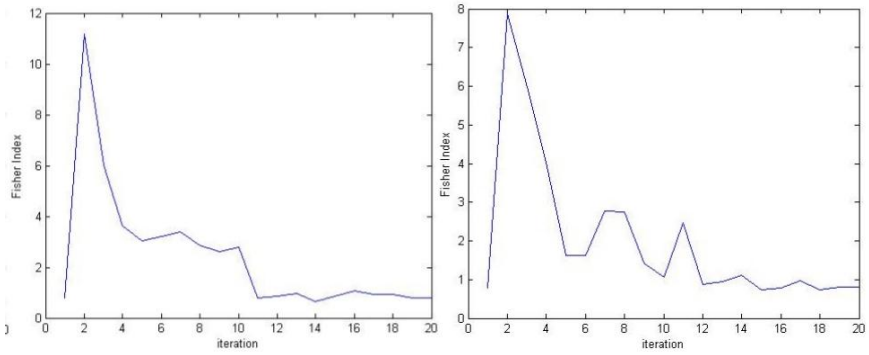
با مشاهده خروجی حاصل از اعمال معیار فیشر بر داده ها با هر دو نوع کدینگ یک و دو (شکلهای ۴,۴ ، ۴,۵ ، ۴,۶ و ۴,۷) در می یابیم که در کدینگ نوع دوم شرایط بهتری برای تصمیم گیری حاکم است، بطوریکه از اعمال این معیار ارزیابی می توان این نتیجه را گرفت که تعداد خوشه مورد نیاز باید بین یازده تا چهارده دسته باشد. لازم به ذکر است که با گرفتن خروجیهای بیشتر از اعمال این الگوریتم بر داده های مسئله ، غالباً به عدد دوازده می رسیم.



شکل ۵,۴ : ارزیابی K-Means با معیار فیشر برای K=1...30 برای داده های با کدینگ نوع یک



شکل ۶,۴ : ارزیابی K-Means با معیار فیشر برای K=1...50 برای داده های با کدینگ نوع دو

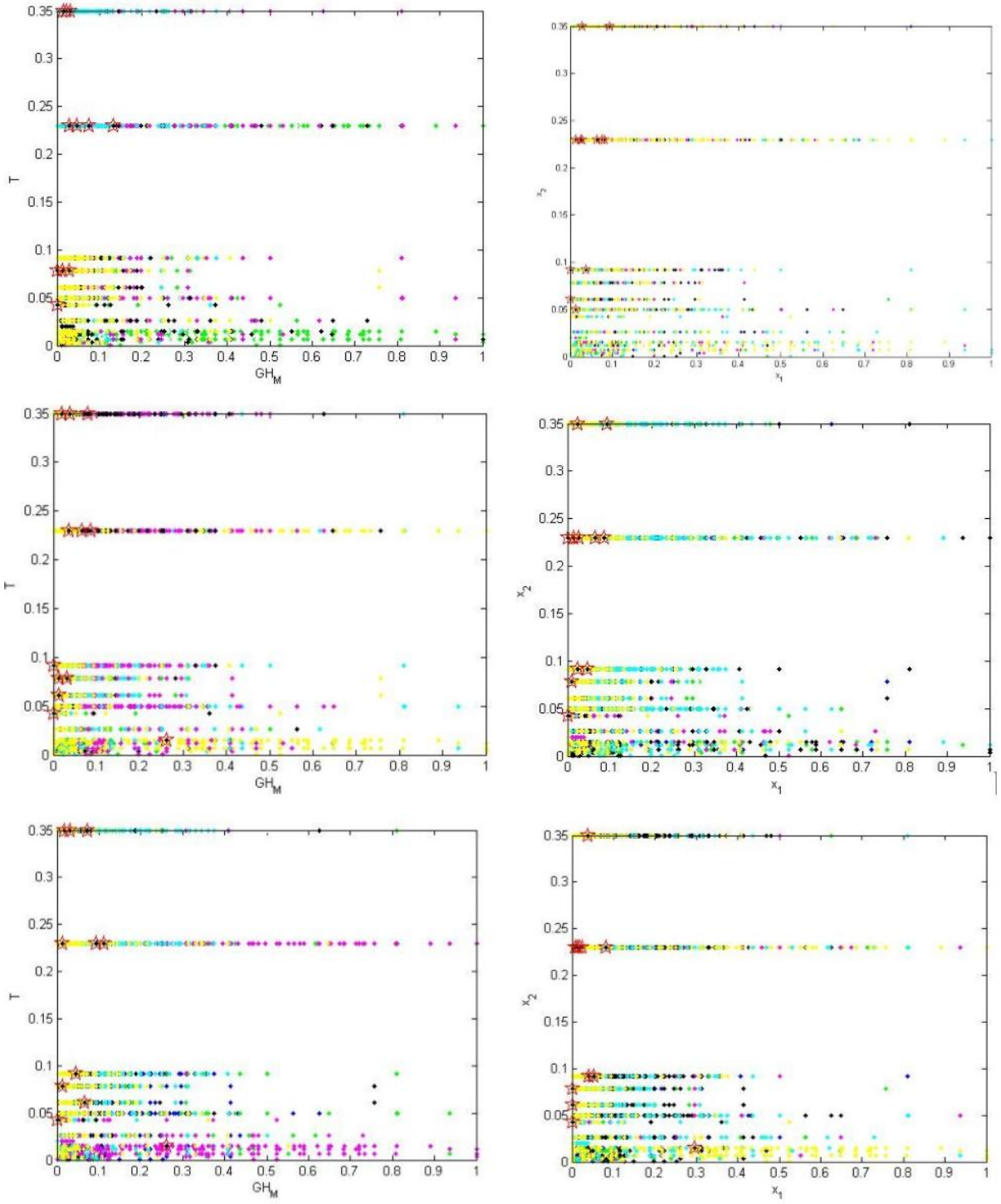


شکل ۷,۴ : ارزیابی K-Means با معیار فیشر برای K=1...30 برای داده های با کدینگ نوع دو

۴-۴- ارزیابی تعداد ویژگیها در نتیجه مساله

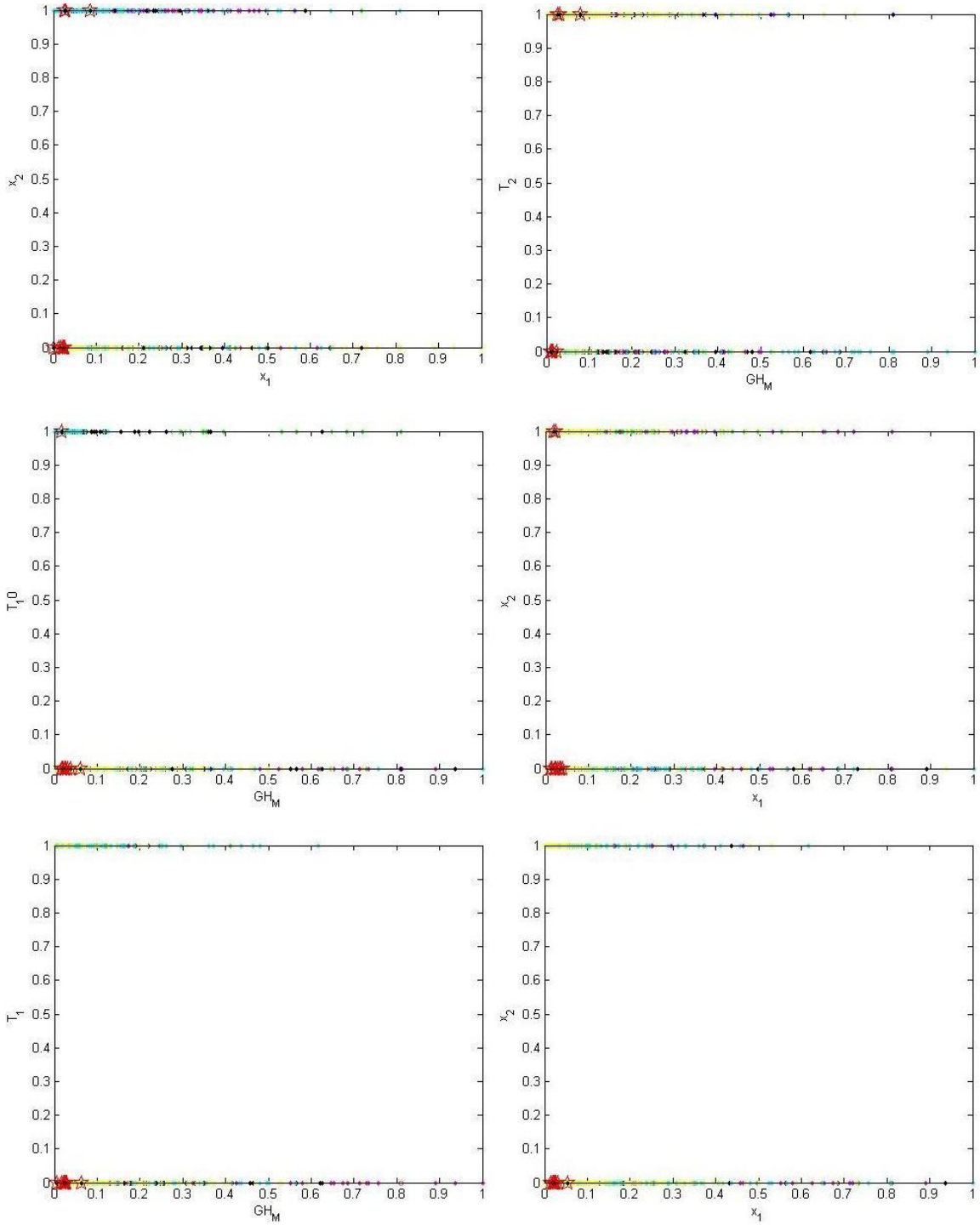
شاید این فرض وجود داشته باشد که انتخاب زیاد پارامترها بعنوان ویژگی و بجهت دخالت دادن آنها در کشف ساختارهای پنهان ممکن است موجب بایاس اولیه الگوریتم به سمت خاصی شود. لذا در یک دوره آزمایش با مقایسه دو ویژگی قیمت منطقه ای و نوع تخلف، این اتفاق بصورت چشمگیری رخ نمی دهد. تَعَدُّدِ خروجی های تهیه شده بدلیل وابستگی K_Means به مقادیر اولیه و همچنین ضعف در مقابله بهینه های محلی صورت پذیرفته است. (شکل ۴,۸ بعنوان نماینده ای از مشاهدات آورده شده مقایسه این کاهش ویژگی در کدینگ نوع دو برای تخلفاتی که بیش از ۵٪ از کل تخلفات را به

خود اختصاص داده اند (تخلفات ردیف های ۱ ، ۲ ، ۳ ، ۴ ، ۱۰ و ۱۶ به ترتیب با حدود ۹,۲۰٪ ، ۳۴,۹۴٪ ، ۷,۸۱٪ ، ۶,۰۹٪ ، ۲۲,۹۸٪ و ۵,۰۰٪) نیز به نتیجه مشابهی می رسیم.



شکل ۸,۴ : ارزیابی مقایسه ای K-Means با ۲۱ ویژگی(سمت راست) و نه ویژگی (سمت چپ) برای داده های با کدینگ نوع یک

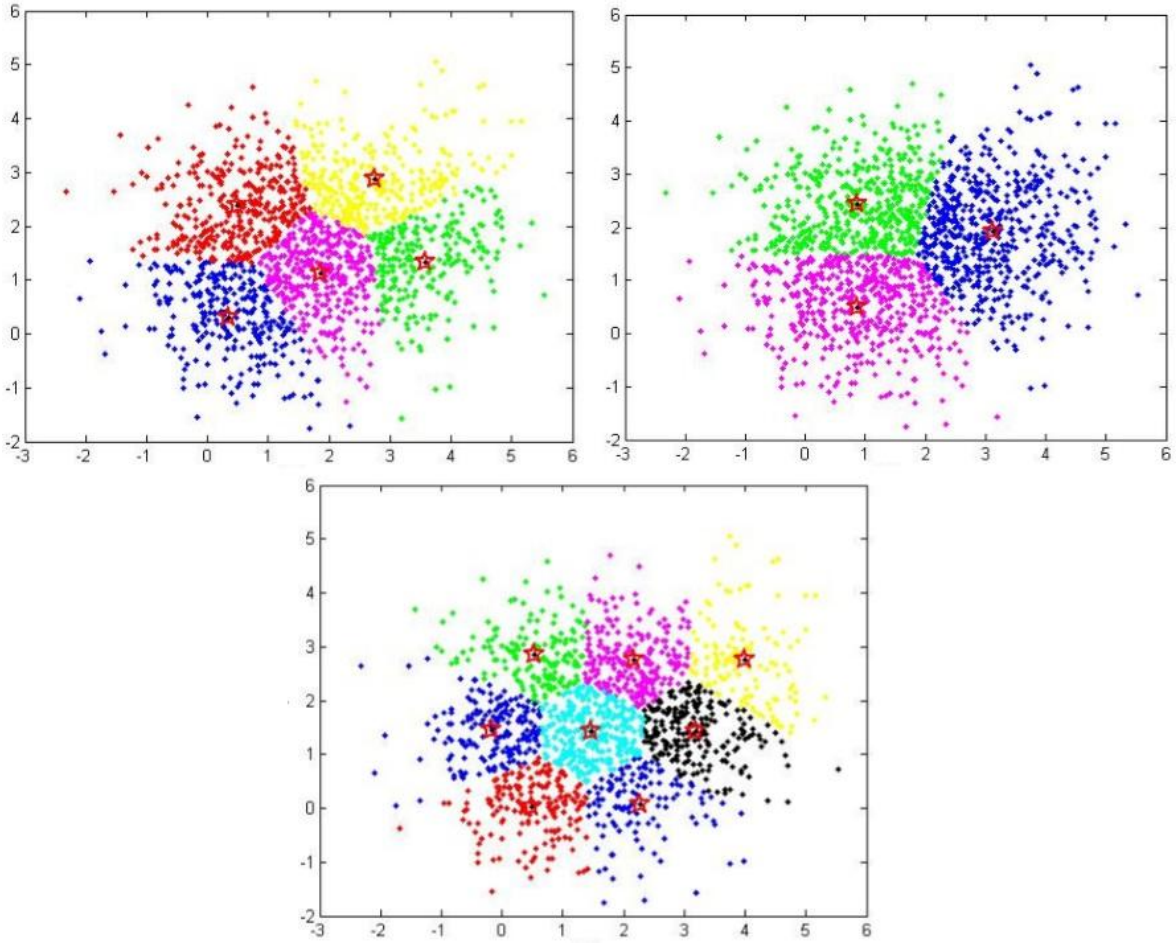
شکل ۴,۹ شهودی از این آزمایش می باشد. (ردیف اول قیمت منطقه ای و تخلف دوم ، ردیف دوم قیمت منطقه ای و تخلف دهم و ردیف سوم قیمت منطقه ای و تخلف اول)



شکل ۹,۴ : ارزیابی مقایسه ای K-Means با ۱۱۲ ویژگی (سمت راست) و ۶۳ ویژگی (سمت چپ) برای داده های با کدینگ نوع دو با معیار فاصله CityBlock

۴-۵- نتایج ناشی از انتخاب **K** های بزرگ یا کوچک در **K-Means** با معیار شباهت فاصله اقلیدوسی

اینکه در انتخابهای غلط K در K-Means چه نتایجی حاصل خواهد شد هم خالی از لطف نیست. برای مشاهده نتایج حاصل از انتخابهای غلط K ، الگوریتم را بر روی داده های نرمال تصادفی ایجاد شده حول سه نقطه X1=(1,1) ، X2=(1,2)، X3=(3,2) و با K=3 ، K=5 و K=8 اجرا می نمائیم.(شکل ۱۰,۴)



شکل ۱۰,۴ : ارزیابی مقایسه ای K-Means با K=3 ، K=5 و K=8

مشاهده می کنید که با افزایش خوشه ها از تعداد واقعی آن، در غالب موارد خوشه ها چون شکل ۷,۲ : تقسیم نمی شوند بلکه تحدید می شوند و خوشه های جدید با اعضائی که به دو یا چند خوشه تعلق داشتند شکل می گیرند. یعنی هر خوشه جدید با یک یا چند خوشه دیگر دارای شباهتهایی خواهد بود. اما در عین حال قطعاً با افزایش K، در هر خوشه میزان شباهتها افزایش خواهد یافت. لذا این موضوع (انتخاب k های بزرگتر) میتواند به حساسیت مسئله برگردد و بر اساس بازنمایی دانش حاصل از ویژگیهای خوشه ها می توان به شباهت های آنها رسید. بعکس اگر تعداد خوشه ها کمتر از تعداد واقعی در نظر گرفته شود، شبهات ها نادیده انگاشته می شوند. و اعضایی که خود یک خوشه تشکیل می دادند در سایر خوشه ها تقسیم می شوند. لذا ریسک خطا در حالتی که K کمتر در نظر گرفته شود، بیشتر از حالتی است که K را بیش از مقدار واقعی در نظر بگیریم.

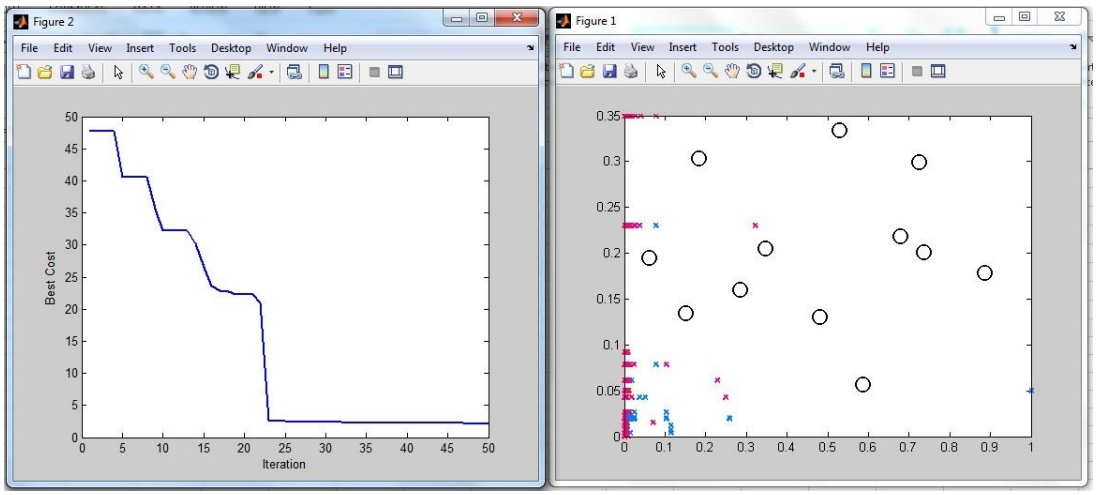
۴-۶- اجرای الگوریتم های فرا ابتکاری اتومات بر روی داده ها

دانستیم که K-Means علی رغم سرعت بسیار بالا و پیاده سازی آسان، دارای معایبی نیز هست. از جمله : حساس بودن به مقادیر اولیه و گرفتار شدن در دم بهینه های محلی، نیاز به مشخص بودن تعداد خوشه ها قبل از اجرا و نداشتن استراتژی گریز از حالتهایی که در یک تکرار تعداد اعضاء یک خوشه صفر می شوند و همچنین نتیجه ای که در بخش اول همین فصل گرفتیم و آن عدم تاثیرقاعده تفاوت حداکثری بین اعضاء خوشه ها علی رغم در نظر گرفتن شباهت حداکثری درون خوشه ها می باشد.

اگر چه مواردی چون عدم امکان خروج از وضعیت بدون عضو شدن یک خوشه با مشاهده و تکرار قابل حل به نظر می رسد اما برای سایر موارد باید لااقل به حدی از اطمینان برسیم. بعنوان نمونه می دانیم که الگوریتمهای ژنتیک و توده ذرات راه حل مناسبی برای گریز از بهینه های محلی هستند. مشکل

انتخاب اولیه تعداد ویژگیها را هم با روشهای اتومات مربوط به این دو الگوریتم می توان آزمایش نمود که درآن صورت با استفاده از شاخص های اعتبار سنجی (شاخص های مورد استفاده DB و CS)، تفاوت حداکثری در بین خوشه ها را هم دنبال کرده ایم.

یافته های پژوهش در خصوص اعمال الگوریتمهای ژنتیک و توده ذرات بر داده های مسئله مشابهت دارد. اگر چه با در نظر گرفتن آستانه ۰,۵ معمولاً تعداد خوشه ها کمتر از ۱۲ دسته در نظر گرفته می شود. البته می توان با تغییر مقدار آستانه به مقادیر کمتر از ۰,۵ (مثلاً ۰,۴ یا ۰,۴۵) میزان اهمیتِ تفاوتِ بین خوشه ها را کمتر نموده و در در خروجی مسئله ۱۲ خوشه مشاهده نمود. که با انجام این تغییر به لحاظ تعداد نتایج با روش K-Means مشابه می باشد. در شکل ۱۱,۴ :یک خروجی حاصل از اعمال الگوریتم اتومات PSO با شاخص DB ، آستانه ۰,۴ ، جمعیت ۱۰۰ و تکرار ۵۰ مرتبه بر داده های با کدینگ نوع دو آمده است. توضیح اینکه شکل سمت چپ نمودار هزینه و شکل سمت راست نمودار دو بعد قیمت منطقه ای و تخلفات می باشد.



شکل ۱۱,۴ : PSO با شاخص DB ، آستانه ۰,۴ ، جمعیت ۱۰۰ و تکرار ۵۰ مرتبه بر داده های با کدینگ نوع اول

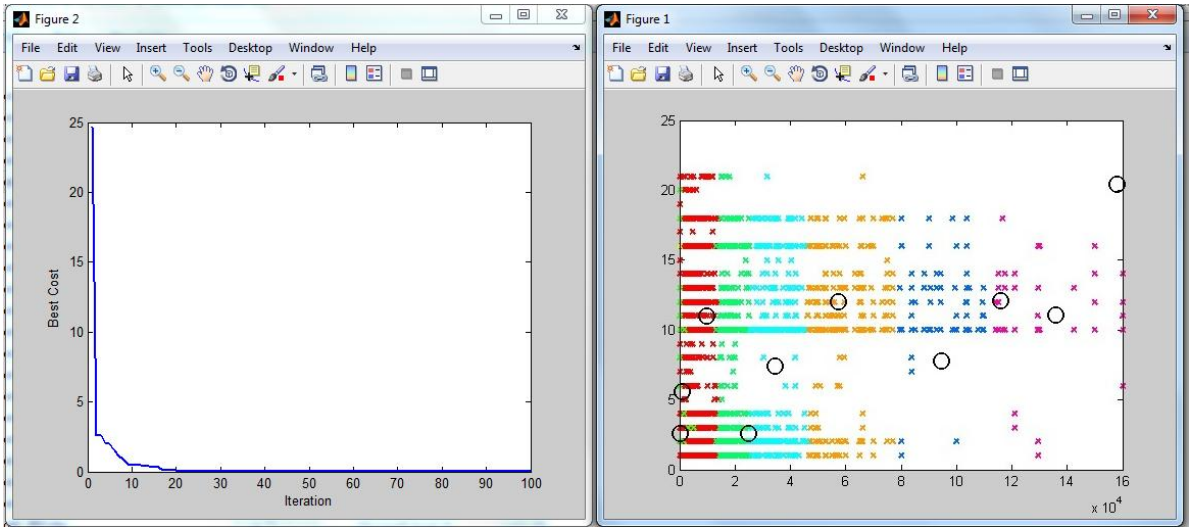
جدول ۵,۴ : مقایسه یک خروجی روشهای GA و PSO اتومات با K-Means (ستون دوم هر یک از بزرگ به کوچک مرتب شده)

GA	
خوشه	فراوانی
10	5224
12	2586
1	1860
8	1727
5	1259
3	1009
6	808
2	607
11	545
7	449
9	363
4	293

K-Means	
خوشه	فراوانی
4	2430
5	2042
2	1938
9	1829
3	1767
6	1731
10	1196
1	1042
11	799
8	697
12	681
7	578

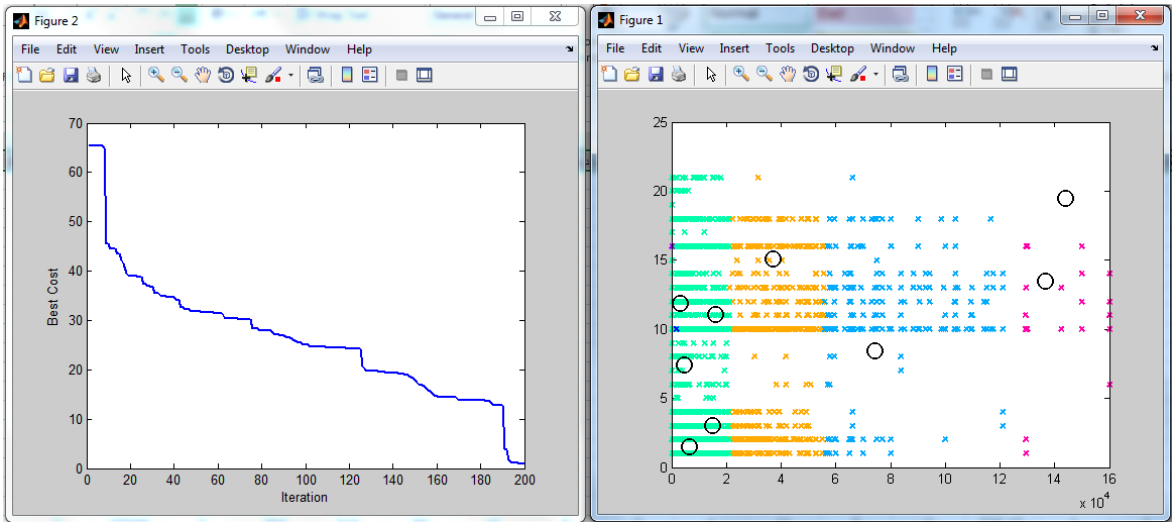
PSO	
خوشه	فراوانی
8	3999
12	3174
1	2039
9	1930
11	1375
2	1365
3	728
4	701
7	510
6	445
10	386
5	78

ملاحظه می شود که تفاوت چندانی در خروجیها به چشم نمی خورد. اگر چه که در تکرارهای متفاوت و خصوصاً در K-Means قطعاً شرایط برای کلیه گروهها متغیر خواهد بود، لیکن با تکرار زیاد بر روی داده های این مسئله خواهیم یافت که تغییرات محدود به بازه هایی است که قابل چشم پوشی بوده است.

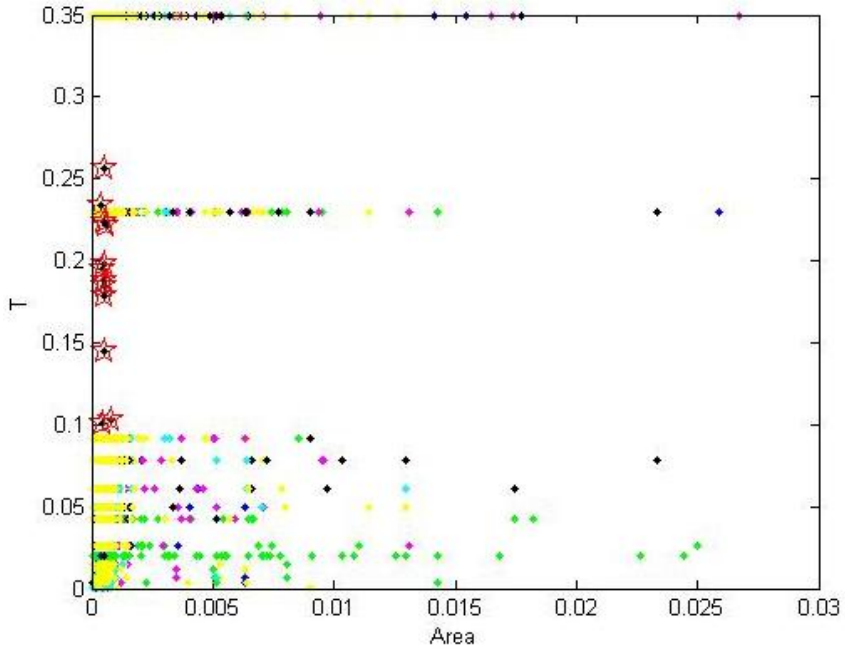


شکل ۱۲,۴ : نمودار مربوط به قیمت منطقه ای و تخلفات در اجرای الگوریتم PSO با k=15 ، شاخص DB=0.5 ، جمعیت ۱۰۰ و تعداد تکرار ۱۰۰ مرتبه بر روی داده های کدینگ نوع اول

یک تحلیل ساده از مشاهده خروجی های شکل ۱۲,۴ :و شکل ۱۳,۴ :آنست که خوشه ها بر اساس قیمت منطقه ای چیدمان یافته اند. یعنی داده های دارای قیمت منطقه ای پائین یک خوشه و به همین ترتیب داده های شامل بالاترین قیمت منطقه ای با تعداد کمترشان به نسبت آنها (خوشه های دارای قیمت منطقه ای پائین) در یک خوشه قرار گرفته اند. (فراموش نکنیم که خوشه بندی را بر اساس ویژگیهای متفاوتی انجام داده ایم) و بنا بر انتظار با افزایش قیمت منطقه ای از میزان تخلفات کاسته می شود. این موضوع با مشاهده نقشه مربوطه و به دو دلیل قابل قبول است. اول آنکه مناطق حاشیه ای در نقشه (دارای قیمت منطقه ای کم) از نظر وسعت سهم بیشتری از املاک دارند و دوم آنکه حاشیه نشینی معمولاً خارج از قواعد، شکل گرفته و توسعه می یابد.



شکل ۱۳,۴ : نمودار مربوط به قیمت منطقه ای و تخلفات در اجرای الگوریتم PSO با k=15 ، شاخص DB=0.4 ، جمعیت ۱۰۰ و تعداد تکرار ۲۰۰ مرتبه بر روی داده های کدینگ نوع دوم (ویژگی قیمت منطقه ای (نرمال نشده) و تخلف ۲ با ضریب ۲۵)



شکل ۱۴,۴ : نمودار مربوط به مساحت و تخلفات در اجرای الگوریتم K-Means با k=12 بر روی داده های کدینگ نوع دوم

مشاهده می نمائید که چنین رابطه ای بین مساحت املاک و تخلفات وجود ندارد. یعنی اینکه نمی توان با افزایش متراژ ساختمان ارتباطی واضح شبیه ارتباط قیمت منطقه ای با خوشه ها یافت.

۴-۷- نتایج اعمال **K-Means** بر داده های مسئله

قبل از تحلیل یافته ها و بر اساس آمار تخلفات(جدول ۶,۴ :) می دانیم که مناطق حاشیه ای از فراوانی قابل توجهی برخوردارند. همچنین انتظار داریم بین وسعت مناطق مسکونی و تعداد تخلفات رابطه مستقیمی وجود داشته باشد.

جدول ۶,۴ : ده منطقه اول شهر بلحاظ تخلفات

منطقه	فراوانی تخلف
9	3072
6	1712
54	1173
46	923
53	679
8	656
7	448
57	445
52	442
42	386

ده تخلف شایع به ترتیب فراوانی در بانک داده مسئله نیز به شرح جدول ۷,۴ :می باشد.

جدول ۷,۴ : ده تخلف شایع

فراوانی تخلف	نوع تخلف[۷۴]
5846	2
3845	10
1540	1
1308	3
1020	4
837	16
710	8
441	18
331	20
256	13

با اطمینان قابل قبول از صحت پیش پردازش ، خوشه بندی و معیار شباهت بکار گرفه شده و در نهایت اعتبار خوشه ها، تحلیل نتایج قطعاً ما را به روبط بین ویژگیها رهنمون خواهد شد. لذا و برای نمونه چند ویژگی (۲۱ تخلف و هشت گونه رای صادره) را با توجه به برخی از دیگر ویژگیهای اعضاء هر خوشه بررسی می نمائیم.

۴-۸- تحلیل خوشه ها و ویژگیهای تخصیص یافته به آنها

سرانجام ماحصل اقدامات پژوهش در تحلیل خروجی رده بندی مشخص خواهد شد. لذا با بررسی ویژگیهای خوشه ها ، روابط موجود را به صورت گزاره های منطقی استخراج می نمائیم.

با نگاهی به جدول ۸,۴ : و همچنین جدول ۹,۴ :به نتایج ذیل می رسیم : (توضیح اینکه فراوانی های بیشتر از ۱۰۰ رنگی شده اند)

[۷۴] - شماره تخلفات بر اساس جدول ۳,۸ آورده شده است.

- خوشه اول : عمدتاً حاوی تخلف ۱۰ و ۱۶ بوده و منجر به صدور رای جریمه گردیده است.
- خوشه دوم : عمدتاً حاوی تخلف ۳ ، ۱ ، ۱۳ و ۲ و شانزدهم بوده و منجر به صدور رای جریمه گردیده است.
- خوشه سوم : عمدتاً حاوی تخلف ۱۰ ، ۳ ، ۴ و ۸ بوده و منجر به صدور رای جریمه گردیده است.
- خوشه چهارم : عمدتاً حاوی تخلف ۲ و ۱۶ بوده و منجر به صدور رای جریمه گردیده است.
- خوشه پنجم : عمدتاً حاوی تخلف ۱۰ بوده و منجر به صدور رای جریمه و اعاده گردیده است.
- خوشه ششم : عمدتاً حاوی تخلف ۲ بوده و منجر به صدور رای جریمه گردیده است.
- خوشه هفتم : عمدتاً حاوی تخلف ۲ بوده و منجر به صدور رای جریمه گردیده است.
- خوشه هشتم : عمدتاً حاوی تخلف ۳ بوده و منجر به صدور رای عوارض گردیده است.
- خوشه نهم : عمدتاً حاوی تخلف ۲ ، ۷ ، ۱ ، ۱۵ بوده و منجر به صدور رای عوارض و رفع تعرض گردیده است.
- خوشه دهم : عمدتاً حاوی تخلف ۲ بوده و منجر به صدور رای جریمه گردیده است.
- خوشه یازدهم : عمدتاً حاوی تخلف ۲ و ۱۰ بوده و منجر به صدور رای جریمه گردیده است.
- خوشه دوازدهم : عمدتاً حاوی تخلف ۲۰ و ۱۸ بوده و منجر به صدور رای جریمه گردیده است.

بدون در نظر گرفتن ارتباط سایر ویژگیها با تخلفات، نسبت آراء نشان دهنده سیاستهای مدیریت شهری در طول زمان در مواجهه با تخلفات ساختمانی یعنی نگاه درآمدی به جریمه های اخذ شده است.

جدول ۸,۴ : توزیع تخلفات در بین خوشه های دوازده گانه استخراج شده از روش K-Means

جمع	T21	T20	T19	T18	T17	T16	T15	T14	T13	T12	T11	T10	T9	T8	T7	T6	T5	T4	T3	T2	T1	خوشه
1042	0	0	0	3	0	213	0	0	8	4	3	694	12	17	0	0	0	59	21	8	0	1
1938	0	0	0	77	5	74	0	12	180	96	20	0	0	0	0	0	0	0	601	149	724	2
1767	2	0	0	0	0	0	0	0	0	0	15	888	0	176	7	56	6	232	301	84	0	3
2430	0	0	0	26	0	314	0	78	21	38	15	0	0	0	0	0	0	0	0	1699	239	4
2042	0	0	0	0	0	3	0	0	0	0	20	2019	0	0	0	0	0	0	0	0	0	5
1731	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1731	0	6
578	0	1	0	0	0	8	0	0	4	0	0	1	0	0	0	0	0	3	0	559	2	7
697	0	0	0	0	0	0	6	0	1	7	0	0	0	0	0	0	0	0	677	0	6	8
1829	1	0	0	24	0	122	7	14	0	51	35	85	0	490	0	0	0	0	304	221	475	9
1196	0	0	0	0	0	37	0	0	27	2	0	1	0	0	0	0	0	20	14	1069	26	10
799	0	4	0	19	1	60	0	11	9	7	6	156	1	27	4	4	0	29	67	320	74	11
681	60	326	3	292	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12

جدول ۹,۴ : توزیع آراء کمیسیون در بین خوشه های دوازده گانه استخراج شده از روش K-Means

خوشه	عوارض	جریمه	تعطیل	اعاده	تامین پارکینگ	تعویق رای	تخریب	رفع تعرض	برائت
1	8	647	0	88	4	1	46	3	143
2	48	1450	0	32	0	0	30	0	209
3	30	1223	2	62	10	0	72	0	198
4	27	1996	2	65	0	0	40	0	127
5	14	1214	1	153	30	0	53	0	322
6	2	1530	0	18	0	0	12	0	51
7	8	502	0	6	1	0	6	0	19
8	31	542	0	0	0	13	0	48	48
9	25	1402	0	34	1	0	89	0	128
10	6	1059	0	20	0	0	5	0	48
11	12	600	0	25	1	0	15	0	78
12	6	380	0	17	0	0	81	0	74

در نگاهی دیگر به خوشه ها و مد نظر قرار دادن ویژگیهای مساحت ، تراکم ، تراکم تجدید نظر ، حد نصاب تفکیک و قیمت منطقه ای، با در نظر گرفتن بیشینه مقدار آن ویژگی در هر خوشه ، میانه و میانگین به ترتیب در ردیفهای یک تا سه برای هر خوش (جدول ۱۰,۴ :) به گزاره های ذیل خواهیم رسید.

جدول ۱۰,۴ : ویژگیهای هفت گانه خوشه های دوازده گانه استخراج شده از روش K-Means

خوشه	مساحت	تعداد طبقات زیر زمین	تعداد طبقات سایر	تراکم	تراکم تجدید نظر	حد نصاب تفکیک	منطقه ای قیمت
1	88321	1	3	1.8	3	1000	150000
	102.7	0	1	1	1.5	200	3960
	326.26	0.17	0.70	1.00	1.59	219.82	8409.61
2	386214.5	2	7	1.8	2.1	2000	100000
	198.35	0	1	1	1.5	200	2904
	426.80	0.22	0.91	0.94	1.31	172.69	6689.48
3	96218	1	3	1.8	3	2000	129500
	186	0	0	1	1.2	200	3960
	350.31	0.20	0.48	0.87	1.28	174.14	8512.05
4	15642	3	7	1.8	3	2000	129500
	190.5	0	0	1	1.2	200	4950
	225.99	0.20	1.10	0.86	1.32	177.46	10617.42
5	10312	1	3	1.8	3	2000	104544
	190	0	0	1	1.2	200	2500
	197.03	0.10	0.52	0.79	1.13	151.64	7068.47
6	124286	2	6	1.8	3	2000	80000
	173	0	1	1	1.5	200	4950
	283.42	0.13	1.26	0.85	1.31	175.79	9553.81
7	100000	3	7	1.8	7.1	3000	83688
	171.655	0	2.5	0.6	2.1	250	9900
	648.21	0.22	2.16	0.71	2.05	246.37	12142.78
8	3657.09	1	3	1.8	3	2000	60000
	198	0	0	1	1	0	2000
	192.04	0.18	0.49	0.78	0.88	91.99	4886.61
9	30000	2	6	1.8	2.1	2000	103620
	120	0	0	1	1.2	200	3960
	218.79	0.16	0.60	0.81	1.00	142.22	6666.15
10	8750	2	6	1.8	3	450	103620
	143.76	0	0	0.6	2.1	250	10850
	178.45	0.14	0.77	0.76	1.88	233.51	14943.21
11	14309	1	5	1.8	2.4	7000	129500
	180	0	0	1	1.2	200	4200
	253.64	0.19	0.89	0.85	1.30	181.70	9876.58
12	40000	1	3	1.8	2.4	2000	160000
	160	0	0	1	1.2	200	2500
	262.89	0.09	0.63	0.85	1.20	171.17	18525.88

- املاک خوشه های ۲ ، ۳ ، ۴ و ۷ بیشترین تعداد زیر زمین را دارا هستند.
- املاک بلند مرتبه به ترتیب در خوشه های ۷ ، ۶ و ۴ قرار دارند.
- بیشترین تراکم به ترتیب مربوط به خوشه های ۱ ، ۲ ، ۳ و ۴ می باشد.
- کمترین تراکم به ترتیب (و از کم به زیاد) به خوشه های ۷ ، ۱۰ ، ۸ و ۵ تعلق دارد.
- بیشترین تراکم تجدید نظر به ترتیب به خوشه های ۷ ، ۱۰ و ۱ تعلق گرفته است.
- کمترین تراکم تجدید نظر به ترتیب (و از کم به زیاد) به خوشه های ۸ ، ۹ و ۵ تعلق دارد.
- بیشترین حد نصاب تفکیک به ترتیب به خوشه های ۷ ، ۱۰ و ۱ تعلق گرفته است.
- کمترین حد نصاب تفکیک به ترتیب (و از کم به زیاد) به خوشه های ۸ ، ۹ و ۵ تعلق دارد.

توجه :

- این نتایج بر اساس شاخص مرکزی میانگین حاصل شده است.
- خوشه های حد وسط در تراکم تجدید نظر و حد نصاب تفکیک یکسان نیستند.

- جداول مربوط به متوسط قیمت منطقه ای خوشه ها و متوسط مساحت املاک خوشه ها :

جدول ۱۱,۴ : متوسط قیمت منطقه ای خوشه ها

خوشه	میانگین
12	18525.88
10	14943.21
7	12142.78
4	10617.42
11	9876.58
6	9553.81
3	8512.05
1	8409.61
5	7068.47
2	6689.48
9	6666.15
8	4886.61

جدول ۱۲,۴ : متوسط مساحت املاک خوشه ها

خوشه	مساحت
7	648.2052
2	426.8004
3	350.3086
1	326.2565
6	283.4215
12	262.8919
11	253.6399
4	225.9897
9	218.7948
5	197.0347
8	192.0431
10	178.4502

جدول ۱۳,۴ : ترکیب خوشه ها از نظر نوع اسکلت

خوشه	بتنی	فلزی	تیر چوبی	آجری	نیمه فلزی	خشت و گل	نیمه بتنی	صنعتی ساز
1	0	0	0	814	158	67	3	0
2	639	0	0	750	362	12	90	0
3	0	0	0	1765	0	0	0	2
4	1410	505	3	0	419	43	30	0
5	0	0	0	2039	1	2	0	0
6	1278	427	9	17	0	0	0	0
7	453	60	0	52	12	0	1	0
8	0	0	0	683	8	6	0	0
9	414	65	0	1039	198	9	38	1
10	893	197	0	39	64	2	0	0
11	279	88	0	315	79	9	25	0
12	0	0	0	0	0	0	681	0

- خوشه های ۴ ، ۶ ، ۱۰ و ۲ بیشترین تعداد ساختمانهای تمام بتنی و تمام فلزی را دارند.
- خوشه های ۱۲ ، ۸ ، ۵ و ۳ فاقد ساختمان تمام بتنی هستند.
- خوشه های ۵ ، ۳ ، ۹ و ۱ بیشترین تعداد املاک آجری را در خود دارند.
- خوشه های ۴ ، ۲ و ۹ بیشترین تعداد املاک نیمه فلزی را در خود دارند.
- خوشه ۱۲ تنها حاوی املاک نیمه بتنی است.

۴-۹- نمونه نتایج نهایی حاصله

- با عنایت به شکل ۱۱,۴ :و شکل ۱۲,۴ :با افزایش قیمت منطقه ای سهم تخلفات کاهش می یابد. البته این موضوع شامل املاک واقع در حریم شهر نمی شود. مطالعه خوشه دوازدهم این

نتیجه را در بر دارد که تجاوز به حریم شهر و ساخت و ساز در حریم، علی رغم قیمت منطقه ای بالا ، اتفاق افتاده است.

- اگر چه منطقی است که املاکی که حد نصاب تفکیک پائین را به خود اختصاص داده اند (خوشه های ۸ و ۹ و ۵) کمترین تراکم تجدید نظر را نیز اخذ نموده اند. اما املاکی که حد نصاب تفکیک بالا را به خود اختصاص داده اند (خوشه های ۷ و ۱۰و ۱) بیشترین تراکم تجدید نظر را نیز اخذ نموده اند. در حالی که بر اساس طرح تفصیلی کمترین تراکم را داشته اند.
- تخلفات رابطه معنا داری با مساحت املاک ندارند. بعنوان مثال در خوشه کم تعداد ۷ بیشترین میانگین مساحت را داریم که می تواند موید کاهش تخلف با افزایش مساحت باشد در حالیکه در خوشه های نسبتاً پر تعداد ۲ ، ۳ و ۱ که در رتبه های بعدی مساحت قرار دارند این قاعده بچشم نمی خورد. همچنین خوشه های ۱۰ ، ۸ و ۵ به ترتیب کمترین میانگین مساحت ها را به خود تخصیص داده اند در حالیکه خوشه های ۱۰ و ۵ نسبتاً پر تعداد هستند. (در نظر داشته باشید که تعداد در یک خوشه نمایانگر تخلف هم هست زیرا هر رکورد مربوط به یک تخلف است)

فصل پنجم

نتایج، پیشنهادات و کارهای آینده

۵-۱- مقدمه

در این پژوهش سعی شد تا با تکیه بر داده کاوی در بانک اطلاعات شهری راهی دیگر جهت تحلیل رفتار شهروندان در حوزه ساخت و ساز و در کنار پژوهش های رایج که عمدتاً مبتنی بر آمار هستند ارائه نمائیم. این روش که جهت تحلیل دلایل وقوع تخلفات ساختمانی با تکیه بر الگوهای پنهان داده های ثبت شده در بانک اطلاعات شهری ارائه گردیده است می تواند در سایر حوزه ها و ساختارهای داده ای نیز بکار گرفته شود. اگر چه که در نگاه اول تصور حل مسئله به شیوه ای جدید مطرح می شود اما در بسیاری از مواقع روشهای هوشمند علاوه بر حل مسئله، ما را به چشم اندازهایی رهنمون می شوند که در نگاه آماری مغفول مانده اند. متخصصان آمار بحق مدعی هستند که در دنیای جدید این آمار هست که گرفتن تصمیمات را برای مدیران ناگزیر می سازد، اما استنتاج آماری مبتنی بر ساختارهای هوشمندانه و یادگیرنده انسانی است که علی رغم محاسن فراوان دارای محدودیتهایی نیز هست. محدودیت هایی چون امکان تصمیم گیری در مواجهه با انبوهی از اطلاعات که اگر چه هر کدام به تنهایی تصمیم مشخصی از انسان هوشمند را رغم می زنند اما در کنار هم به یک نتیجه جامع خواهند رسید که شاید با نتایج حاصل از تحلیل تک تک اطلاعات، همسو نباشد.

این پژوهش شامل دو طیف از نتایج می باشد. اول ارائه یک روش مبتنی بر داده کاوی بدون ناظر (با استفاده از رده بندی داده ها) و دستیابی به دانش جهت تصمیم گیری مدیریت شهری، و دوم ارزیابی رویکردها در مواجهه با داده های مختلط یا ترکیبی. در ادامه به ارائه خلاصه نتایج، پیشنهاداتی برای شهرداری شاهرود و ذکر کارهای مفید برای آینده می پردازیم.

۵-۲- خلاصه نتایج حاصله

- فاقد ارزش دانستن خوشه بندی با معیارهای شباهت متریک بخاطر شکاف و فاصله غیر قابل قبول بین معیارهای شباهت متریک برای انواع داده های طبقه بندی و داده های عددی، در این پژوهش مورد نقد قرار گرفت. با عنایت به پاسخهای ارزشمند و منطقی حاصله می توان چالش پیش روی این تفکر را در نظر نگرفتن پیش پردازش داده ها و نرمال سازی آنها بر اساس خواصشان دانست. بعبارت ساده تر و بعنوان نمونه وقتی نرمال سازی یک داده طبقه بندی بر اساس فرمولهای مبتنی بر فراوانی شکل می گیرد، اعداد حاصل حاوی اطلاعات عددی مرتبط با فراوانی ویژگی خواهند بود.
- همچنین در مقایسه دو روش کدینگ بکار گرفته شده در مسئله به نظر می رسد که علی رغم موثر بودن روش کدینگ نوع دوم در بسیاری از مواقع بدلیل نرمال سازی که در بند قبل اشاره شد، تفاوت چشمگیری در نتایج حاصله وجود نداشته باشد.
- با عنایت به اینکه با افزایش قیمت منطقه ای عموماً سهم تخلفات کاهش می یابد، به نظر می رسد که در دید کلان، این گزینه عامل بازدارنده موثری خواهد بود.
- املاکی که حد نصاب تفکیک بالا را به خود اختصاص داده اند بیشترین تراکم تجدید نظر را نیز اخذ نموده اند. در حالی که بر اساس طرح تفصیلی کمترین تراکم را داشته اند. لذا تجدید نظر در طرح تفصیلی این املاک ضروری به نظر می رسد.
- برخلاف تصور اولیه، تخلفات رابطه معنا داری با مساحت املاک ندارند و بیش از آنکه مساحت املاک در وقوع تخلف موثر باشد سایر ویژگیهای آن از جمله موقعیت مکانی ملک در وقوع تخلفات موثر هستند.

۵-۳- پیشنهادات یا نقطه نظرهای پژوهشگر برای شهرداری شاهرود

برخی نتایج پژوهش از قبیل اصلاح روال انجام کار و پیشنهادات نرم افزاری برای شهرداری شاهرود به شرح ذیل می باشد:

۵-۳-۱- اصلاحات عملکردی چون

- ضرورت تکمیل و ورود اطلاعات برخی فیلدهای ارزشمند در هر پرونده
  - از جمله اطلاعات کامل مالک

۵-۳-۲- اصلاحات نرم افزاری چون

- خودکار سازی محاسبات عوارض در بخشهای مختلف از قبیل کسب و پیشه ، نوسازی و شهرسازی، زیرا در غیر اینصورت علاوه بر پذیرش ریسک خطاهای سهوی ، امکان انجام خطا را برای پرسنل فراهم می نماید.
- خودکار سازی برخی قواعد
  - بعنوان مثال آراء بدوی کمیسیون ماده صد با گذشت مدت مقرر قانونی و عدم اعتراض به رای (از سوی مالک یا شهرداری) خود به خود قطعی خواهند شد در حالی که در سیستم تغییر وضعیت نمی یابند، لذا باید پس از گذشت مدت قانونی بصورت خودکار به قطعی تبدیل گردند.
- امکان ورود اطلاعات با فرمتهای متفاوت و تکیه نرم افزار بر مطالب توضیحی کاربران نتایج ذیل را در بر دارد :
  - دشوار نمودن ورود اطلاعات ایجاد سختی کار برای کاربران
  - عدم امکان گرفتن گزارشات آماری به صورت مقایسه ای

- عدم شفافیت عملکرد کاربران
- عدم امکان نظارت بر کاربران
- عدم وجود وحدت رویه

لذا اصلاح نرم افزاری برای جایگزین نمودن گزینه های انتخابی بجای توضیحات و همچنین چک نمودن برخی مغایرتها قبل از ثبت اطلاعات ضروری می باشد.

۵-۳-۳- اصلاحات فرآیندی

- مطالعه مجدد و مورد به مورد فرآیندها و اصلاح آنها و اعمال تغییرات همزمان در نرم افزارها
- بعنوان مثال جایگزینی لینک بین داده ها بجای ورود اطلاعات مشابه و غیر ضرور

۵-۴- پیشنهاد برای پژوهش های آتی

۵-۴-۱- در این پژوهش بر اساس محدودیت ها و شرایط پایگاه داده و چگونگی ثبت اطلاعات ، سیاستهایی اعمال گردید تا با اصلاح ساختار رکورد های داده، آن محدودیتها را مرتفع نمائیم. اما به نظر می رسد که در بسیاری از مواقع اعمال محدودیتهایی در خوشه بندی مورد نیاز می باشد. اینکار می تواند تلفیقی از یادگیری با ناظر و یادگیری بدون ناظر باشد. بعنوان مثال خوشه بندی داده ها بنحوی صورت پذیرد که خوشه ها با محدودیت تعداد روبرو باشند. یا اینکه یک یا چند مرکز بصورت حدودی مشخص شده باشند. و یا آنکه تاثیر یک یا چند ویژگی در فرآیند رده بندی بیشتر و یا کمتر از سایر ویژگیها در نظر گرفته شود. بنابراین مطالعه چگونگی اعمال محدودیتها بر الگوریتم های خوشه بندی بعنوان پژوهش آتی مفید بنظر می رسد.

۵-۴-۲- به عنوان کار آینده ارائه روشهای مبتنی بر کمی سازی داده های کیفی (داده های اسمی) با حفظ ارزشهایشان بعنوان یک پژوهش ارزشمند پیشنهاد می شود .

فهرست منابع

۱. کاربرد رویکرد شکل مبنا در تدوین ضوابط و مقررات شهرسازی و آیین نامه منطقه بندی. اعتصامی پور, محسن. ۱۳۹۱.

۲. تبیین وضعیت و شناخت عوامل مؤثر بر تخلفات ساختمانی در کلانشهرهای ایران مطالعهٔ موردی مناطق پانزده گانهٔ شهر اصفهان. محمدی, جمال. ۱۳۹۳.

۳. بررسی انگیزه های تخلف احداث بنای مازاد بر تراکم ساختمانی در شهر تهران. سرخیلی, الناز. ۱۳۹۱.

۴. بررسی و تحلیل عوامل مؤثر بر تخلفات ساختمانی در شهرهای کوچک نمونهٔ موردی شهر بابلسر. پژوهان, موسی. ۱۳۹۲.

۵. موانع اجرایی ضوابط شهرسازی و ارائه راهکارهای مناسب در جهت جلوگیری از تخلفات ساختمانی مطالعه موردی کلان شهر تبریز. ظاهری, محمد. ۱۳۸۵.

۶. بررسی علل تخلفات ساختمانی در شهرهای استان لرستان. بهاروندی, عطاالدین. ۱۳۹۲.

۷. بررسی تأثیر تصمیمات کمیسیون ماده ۱۰۰ شهرداری در کنترل تخلفات ساختمانی. بهمنی منفرد, هادی. ۱۳۹۱.

8. Automatic Clustering of Mixed Data Using Genetic Algorithm. M.Yaghini, M.Vard. September 2012.

9. Clustering of Data with Mixed Attributes based. M.Soundaryadevi, L.S.Jayashree. March 2014.

10. Unsupervised learning with mixed numeric and nominal data. C. Li, G. Biswas. 2002.

11. A Two-Step Method for Clustering Mixed Categroical and Numeric Data. Ming-Yi Shih, Jar-Wen Jheng and Lien-Fu Lai.

12. Spatial data mining implementation Alternatives and performances. Zeitouni, Nadjim.

13. IMPROVING DATA INTEGRATION FOR DATA WAREHOUSE:A DATA MINING APPROACH. Kaloyanova, Kalinka Mihaylova. 2005.

14. A.K. JAIN, M.N. MURTY and P.J. FLYNN. Data Clustering: A Review. 1999.

15. Introduction to Machine Learning. E, Alpaydin. 2004.

16. Clustering. Keller, F.

17. Principles of Knowledge Discovery in Data : Clustering I. J, Sander. Alberta : Department of Computing Science University of Alberta, 2003.

18. Pattern Classification And Scene Analysis. R. O. Duda, P. E. Hart, D. G. Stork. 2000.

19. Statistical Pattern Recognition. R, A. 2002.

20. A Review of Clustering Algorithms as Applied in IR. He, Q. 1999.

21. Spoken Language Processing. X. Huang, A. Acero, H. W. Hon. 2000.

22. Cluster Validity Measurement Techniques. F. Kovács, C. Legány, A. Babos. Budapest : University of Technology and Economics, 2003.

23. Automatic Clustering Using an Improved Differential Evolution Algorithm. Swagatam Das, Ajith Abraham.

۲۴. کلامی هریس, سیدمصطفی. الگوریتم های فرا ابتکاری - مبانی نظری و پیاده سازی در متلب.

25. Scalable Algorithms for Clustering Large Datasets with Mixed Type Attributes. He, Z., Xu, X., Deng, S. s.l. : International Journal of Intelligent Systems, 2005, Vol. 20 No. 10.

26. CLUSTERING LARGE DATA SETS WITH MIXED NUMERIC AND CATEGORICAL VALUE. HUANG, ZHEXUE.

27. Clustering Mixed Data Based on Evidence Accumulation. Luo, H., Kong, F,Li. 2006.

28. Cluster Analysis for Applications. Anderberg, M. R. New York : s.n., 1973.

29. Enhancing k-means algorithm with initial cluster centers derived from data partitioning along the data axis with the highest variance. S., Auwatanamongkol. 2007.

30. Data Mining-Practical Machine Learning Tools And Techniques. Witten, L.H., Frank, E. 2005.

Abstract

Major issues of cities including chaos and disorder are attributed to quantitative construction violations, Investigating the reasons behind this violations and finding the relationship between the violations and the influencing parameters promote the development of the cities.

The study aims at finding the hidden pattern of urban constraction data bases the discover the relationships between these variables and the construction violations.

Data gathering and analyzing were the most significant and the most time-consuming parts of the project with preserving the values of the data being the main priority.

In the study methods and approaches of other researchers concerning combined data , the importance of selecting and finding the feature extraction, preparations and pre-procen including techniques of purifying , integrating , data reduction and data conversion have all been investigated.

And while assessing the k-mains algorithm an data , the significance of Similarity measures and validity of clusters along with comparison of research out put with the output of automatic PSO and GA algorithms have been considered.

Considering the correct choice of the number of clusters and assessing the impact of the number of variables in the result , the research result has been presented in the form of a comparative table with organizing the central feature of each cluster on logical propositions.

The study ends with presenting the relationship between the regional price of estate and their area or the relationship between constraction in privancy an unauthorized constraction.

Keyword : data mining, mixed data, construction violations

E-learning Center, University of Shahrood

Faculty Computer Engineering

Using Data Mining Technique in the Clustering Unauthorized Construction And Analysis of Reasons Abuse

Davood Mazjee

Supervisor :

Dr.Hamid Hasanpoor

February 2016