چکیده

با گسترش سیستم‌های پایگاهی و حجم بالای داده‌های ذخیره شده در آن‌ها، به ابزاری نیازا ست تا بتوان این داده‌ها را پردازش کرد و اطلاعات حا صل از آن را در اختیار کاربران قرار داد. در واقع امروزه به روش‌هایی نیازداریم که با کمترین دخالت کاربر و به صورت خودکار، الگوها و رابطه‌های منطقی را بیان نمایند. یکی از روش‌های بسیار مهمی که با آن می توان الگوهای مفیدی را در میان داده‌ها تشخیص داد، داده‌کاوی است. خوشه‌بندی یکی از مهم ترین روش‌های داده‌کاوی است. خوشه‌بندی در مسائل گوناگونی در زمینه‌های مختلف از جمله تشخیص الگو، مهندسی، پردازش‌تصویر و غیره مورد استفاده قرار می‌گیرد. تاکنون الگوریتم‌های مختلفی برای انجام خوشه‌بندی ارائه شده است. در این پژوهش به معرفی چند روش کارآمد در زمینه مسائل خوشه‌بندی می‌پردازیم. این روش‌ها عبارتند از :

الف) الگوریتم ژنتیک گروه‌بند

ب) الگوریتم ژنتیک با بازآرایی ژن خوشه‌بندی

الگوریتم ژنتیک گروه‌بند برگرفته از الگوریتم ژنتیک کلاسیک است که برای مسائل خوشه‌بندی توصیف می شود و مدل قبلی را بهبود می‌بخشد. در ادامه اجرای الگوریتم مطالعه می‌شود و تجارب به دست آمده از آزمون روی یکسری داده و نحوی مطالعه بر روی الگوریتم گروه‌بندی برای مسائل خوشه‌بندی را کامل می‌کند. در پایان یک الگوریتم پیشنهادی جهت خوشه‌بندی که کارایی بهتری نسبت به بعضی الگوریتم‌های معرفی شده دارد، ارائه می‌گردد.

فهرست مطالب

فهرست شکل ها

فهرست جداول

مقدمه کلی

داده‌ها و الگو یکی از شاخص‌های مهم در عصر اطلاعات می باشند و همگام با پیشرفت در زمینه طبقه‌بندی داده‌ها و استخراج آن‌ها، روش‌های دستی تحلیل داده‌ها از رده خارج شده و استفاده از روش‌های کامپیوتری تحلیل معمول و غیرقابل انکار گردیده‌اند.

بشر به طورذاتی با مواجه‌شدن با مسائل پیچیده شروع به طبقه‌بندی مشاهدات با استفاده از ویژگی‌های آن‌ها که درعلوم مختلف کاربرد دارد می‌کند. برای رسیدن به این منظور، اولین کار موردنظر انسان طبقه‌بندی داده‌های جمع آوری شده و استخراج گروه‌های نمایانگر با خصوصیتی از سیستم مورد مطالعه می‌باشد. همچنین انسان‌ها بیشتر مواقع به استفاده از روش‌هایی برای نمایش روابط موجود در سیستم‌های پیچیده می پردازند. نمایش روابط موجود در سیستم‌های پیچیده با داشتن چندین ویژگی از آن‌ها بسیار مشکل می باشد. خوشه‌بندی از جمله روش‌های پرکاربرد در تجزیه و تحلیل داده‌ها است.

خوشه‌بندی شامل روش‌های بسیار متنوعی است که علیرغم کاربرد در بسیاری از علوم، روش‌ها از نظر هدف، الگوریتم و نتایج با یکدیگر یکسان نیستند. مقالات وکتاب‌های زیادی در ارتباط با روش‌های خوشه‌بندی نوشته شده است، اما هنوز در انتخاب صحیح روش خوشه‌بندی ملاک روشنی وجود ندارد و عدم انتخاب روش صحیح از بین روش‌های موجود، استخراج نادرست الگوها را موجب می‌گردد. خوشه‌بندی، تقسیم‌بندی الگوریتم‌های خوشه‌بندی، طبقه‌بندی، مراحل خوشه‌بندی و معیارهای ارزیابی خوشه‌بندی و الگوریتم ژنتیک در فصل ۱ به طور مفصل تشریح شده است. یکی دیگر از مسایلی که در این پایان‌نامه مورد بررسی می‌گیرد ، مسائل گروه‌بندی است. الگوریتم ژنتیک برای مسائل گروه‌بندی مشابه الگوریتم ژنتیک معمولی است و تنها نحوه کدگذاری آن به شکلی است که برای مسائل گروه‌بندی مناسب می‌باشد. با استفاده از الگوریتم ژنتیک برای مسائل گروه‌بندی، امانوئل فالکنایر[1] متوجه شد که یکسری از اشکالات قابل توجه در الگوریتم ژنتیک برای مسائل گروه‌بندی وجود دارد.

[1]Emanuel Falkenauer

فالکنایر در مقاله‌اش با عنوان مسائل گروه‌بندی و الگوریتم‌های ژنتیک[۱] وارد جزئیاتی از نقص‌های ناشی از اعمال الگوریتم ژنتیک کلاسیک بر روی مسائل گروه‌بندی، شده است. به بیان او کدگذاری بدون واسطه، اثر بسیار نامطلوبی برروی اطلاعات دارد. همچنین فرآیند ترکیب یا تقاطع یک کروموزوم می تواند به راحتی به راه‌حلی برسد که از هیچ یک از صفات والدین استفاده نمی‌کند. به طور م شابه م سائلی در رابطه با به کاربردن طرح کدگذاری ا ستاندارد و عملگر جهش وجود دارد. در ادامه الگوریتم ژنتیک گروه‌بند برای مسائل خوشه‌بندی معرفی و سپس به آن پرداخته می شود. در ادامه رفتار دو الگوریتم k - میانگین و خوشه‌بندی مبتنی بر چگالی را با الگوریتم GGA مقایسه می‌کنیم. نتایج و آزمایشات در انتهای فصل ۲ آمده است.

در ف صل ۳ ضمن ارائه تعاریف مورد نیاز، الگویتم ژنتیک برای خو شه بندی با روش k- میانگین با استفاده از بازآرایی ژن معرفی می‌گردد. سپس به منظور آزمون عملکرد الگوریتم خوشه‌بندی با استفاده از بازآرایی ژن، آزمایش‌ها بر روی داده‌ها در دنیای واقعی از UCI تصاویر سنجش از راه دور، انجام و نتایج نشان داده خواهد شد. نهایتا در این بخش، اجرای خوشه‌بندی با GAGR، خوشه‌بندی با GA، خوشه‌بندی با KGA و خوشه‌بندی با الگوریتم k- میانگین با آزمایشاتی که بر مبنای شش مجموعه داده‌ها در دنیای واقعی است، مقایسه می‌گردند. سپس روشی جدید برای بهبود الگوریتم ژنتیک گروه‌بند ارائه و نتایج وآزمایشات که بیانگر بهبود الگوریتم هستند، ارائه می‌گردد.

فصل اول
مقدمات و پیش‌نیازها

۱-۱-مقدمه

جامعه‌ی مبتنی بر اطلاعات را می‌توان به عنوان جامعه‌ای تعریف نمود که بخش غالب اجتماع به جای کارهای فیزیکی در گیرکارهای فکری هستند. در چنین جامعه‌ای بیشترین توجه به فعالیت‌های اطلاعاتی از قبیل: فراهم‌آوری، پردازش، تولید، ثبت، انتقال، اشاعه و مدیریت اطلاعات مبذول می‌گردد وبیشترین هزینه‌ها صرف فرایندهای اطلاعاتی می‌شود.

با گسترش سیستم‌های پایگاهی و حجم بالای داده‌های ذخیره شده در این سیستم‌ها، به ابزاری نیازاست تا بتوان این داده‌ها را پردازش کرد و اطلاعات حاصل از آن را در اختیار کاربران قرار داد. در واقع امروزه به روش‌هایی نیاز داریم که به اصطلاح به کشف دانش[۱] بپردازند. یعنی روش‌هایی که با کمترین دخالت کاربر و به صورت خودکار، الگوها و رابطه‌های منطقی را بیان نمایند. یکی از روش های بسیار مهمی که با آن می توان الگوهای مفیدی را در میان داده‌ها تشخیص داد، داده‌کاوی است[۲].

۱-۱-۱ داده‌کاوی

از روش‌های بسیار مهم تشخیص الگوهای مفید در میان داده‌ها، داده‌کاوی است. این روش که با حداقل دخالت کاربران همراه است اطلاعاتی را در اختیار آن‌ها و تحلیل‌گران قرار می‌دهد تا براساس آن‌ها تصمیمات مهم و حیاتی در سازمانشان اتخاذ نمایند.

باید توجه داشت که اصطلاح داده‌کاوی زمانی به کار برده می‌شود که با حجم بزرگی از داده‌ها، در حد گیگا یا ترابایت، مواجه باشیم. در تمامی منابع داده‌کاوی بر این مطلب تاکید شده است. هرچه حجم داده‌ها بیشتر و روابط میان آن‌ها پیچیده‌تر باشد، دسترسی به اطلاعات نهفته در میان داده‌ها مشکل‌تر می شود و نقش داده‌کاوی به عنوان یکی از روش‌های کشف دانش، آشکارتر می‌گردد[۳].

[۱] Knowledge Discovery

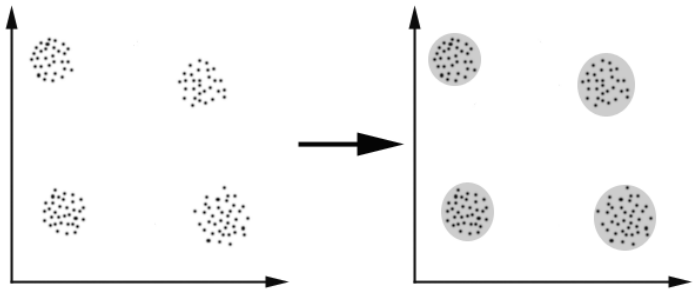
۱-۲- خوشه‌بندی و طبقه بندی

۱-۲-۱ خوشه‌بندی

خوشه‌بندی از مهم‌ترین روش‌های داده‌کاوی و یک زیرگروه مهم از تکنیک‌های یادگیری ماشین بدون نظارت است که شامل گروه‌بندی داده‌ها به خوشه‌های جدا از هم است. به صورت کلی به گروه‌بندی اشیاء داده‌ای به گروه‌های مجزا (خوشه)، خوشه‌بندی می‌گویند. خوشه‌بندی در مسائل گوناگونی در زمینه‌های مختلف از جمله تشخیص الگو، مهندسی، پردازش‌تصویر وغیره مورد استفاده قرار می‌گیرد. دلیل اصلی استفاده از خوشه بندی این است که بعضی مواقع اطلاعات خاصی در مورد ویژگی‌ها و کیفیت خوشه‌ها به علت پیچیدگی پایگاه داده در دسترس نمی‌باشد. در نتیجه بایستی از روشی که توانایی یافتن الگوها از فضای جستجو را دارد، استفاده کنیم[۵،۴].

در تعریفی دیگر به سازماندهی‌داده‌ها در خوشه‌ها، تحلیل خوشه بندی می‌گویند. تحلیل خوشه‌بندی به معنی سازماندهی الگوهای جمع آوری شده است. سازماندهی و تحلیل بر این مبنا صورت می گیرد که داده‌های هر خوشه حداکثر درجه شباهت و داده‌های متعلق به خوشه‌های مختلف دارای حداکثر درجه عدم شباهت باشند. برای مشابه بودن می توان معیارهای مختلفی را در نظر گرفت. مثلا می‌توان معیار فاصله را برای خوشه‌بندی مورد استفاده قرار داد و اشیائی را که به یکدیگر نزدیک‌تر هستند را به عنوان یک خوشه در نظر گرفت که به این نوع خوشه‌بندی، خوشه‌بندی مبتنی بر فاصله[۱] نیز گفته می شود. به عنوان مثال در شکل ۱-۱ نمونه‌های ورودی در سمت چپ به چهار خوشه مشابه شکل سمت راست تقسیم می شوند. در این مثال هر یک از نمونه‌های ورودی به یکی از خوشه‌ها تعلق دارد و نمونه ای وجود ندارد که متعلق به بیش از یک خوشه باشد.

[۱] Distance-based Clustering



شکل۱-۱: خوشه‌بندی نمونه‌های ورودی

هر داده نشان دهنده برداری از مقادیر کمی در یک فضای چند بعدی است.

روش‌های تحلیل به دو گروه عمده سلسله مراتبی و تفکیکی تقسیم می‌گردند:

۱. خوشه‌بندی سلسله مراتبی عبارت است از تشکیل متوالی گروه‌هایی که اعضای آنها بیشترین شباهت را باهم دارند ، یا جداسازی متوالی گروه‌هایی که اعضای آنها دارای بیشترین اختلاف با هم هستند. در این نوع خوشه بندی مشکلات ناشی از مقداردهی اولیه و کمینه‌های محلی وجود ندارد.

۲. خوشه بندی تفکیکی، شامل به دست‌آوردن تفکیکی ازداده‌های ورودی در تعداد مشخصی از خوشه‌ها می‌باشد. چنین روش‌هایی معمولاً به دنبال جداسازی هستند که تابع شایستگی را به صورت محلی بهینه کند. برای بهبود کیفیت خوشه‌بندی الگوریتم چندین بار و در نقاط شروع مختلف اجرا می‌شود و بهترین وضعیت بدست‌آمده از کل دفعات اجرا به عنوان خروجی خوشه‌بندی انتخاب می‌گردد[۷،۶].

۱-۲-۲ تقسیم‌بندی الگوریتم‌های خوشه‌بندی

روش‌های مختلفی برای تقسیم بندی الگوریتم‌های خوشه‌بندی وجود دارد، اما در اینجا با یک دیدگاه متفاوت آن‌ها را به چهار گروه تقسیم می‌کنیم:

۱. د سته اول یک مفهوم کلی از خو شه‌بندی را بر ا ساس این نظر که داده‌های نزدیک به هم باید در یک خو شه قرار گیرند به کار می‌برد. به عنوان نمونه از الگوریتم‌هایی که بر ا ساس

این اصـــل پیاده‌ســـازی می‌شـــوند می‌توان موارد زیر را نام برد: روش‌های بر پایه‌ی چگالی [۱۰،۹،۸]، روش‌های نزدیکترین هم سایه[۱۱]، خو شه‌بندی تراکمی ات صال تک[۱۲]. این روش‌ها به راحتی هر شـــکلی از خوشـــه‌بندی را انجام می‌دهند، اما وقتی فاصـــله مجزا بین خوشه‌ها کم است، در تعیین خوشه‌ها دچار مشکل می‌شوند.

۲. د سته دوم با توجه به متغیر داخل خو شه‌ها، جواب نهایی را به د ست می آورند. این د سته شامل الگوریتم‌هایی مثل k- میانگین [۱۴،۱۳]، خو شه‌بندی یادگیری شبکه [۱۶،۱۵] و خو شه‌بندی بر ا ساس مدل[۱۷] هستند. این روش‌ها برای خو شه‌های کاملا مجزا و کروی موثرتر ه ستند، اما برای خو شه‌هایی با ساختار پیچیده‌تر، ممکن ا ست منجر به شکست شوند.

۳. دسته سوم خوشه‌بندی سطر و ستون ماتریس را به صورت همزمان انجام می‌دهند. از این نوع می‌توان الگوریتم‌های خوشه‌بندی دوگانه [۱۸] را نام برد. هدف این روش این است که با اجرای خوشـــه‌بندی همزمان ســـطر وســـتون ماتریس‌ها، به جای خوشـــه‌بندی مجزا برای ســـطر و ســـتون، زیرگروه‌هایی از ســـطرها و زیرگروه‌هایی از ســـتون‌ها را بشـــناســـند. پس خوشـــه‌بندی دوگانه مدل‌های محلی را تولید می‌کند، در حالی‌که این روش خوشـــه‌بندی مدل‌های کلی را محا سبه می‌کند. خو شه‌های معرفی شده تو سط این الگوریتم‌ها دو به دو ناسازگار یا جامع نیستند. یک داده ممکن است متعلق به هیچ خوشه‌ای نباشد یا در بیشتر از یک خوشه باشد.

۴. دســـته چهارم ویژگی‌های مختلف مجموعه داده را بهینه می‌کند. این الگوریتم‌ها می‌توانند به خو شه‌بندی د سته جمعی تقسیم شوند[۱۹] که نتیجه را به یک موردی با کیفیت بهتر ترکیب می‌کنند و روش‌های خوشـــه‌بندی چند منظوره [۱۸] که یک تخمینی از کیفیت همه حل‌های خوشـــه‌بندی فراهم می‌کنند. اگرچه اغلب الگوریتم‌ها قوی هســـتند اما بدون اشکال هم نیستند.

یک خو شه‌بندی خوب باید هم تو سط الگوریتم تولید شده و هم یک معیار خارجی تایید شود. از بین الگوریتم‌های دســـته دوم، الگوریتم k- میانگین کاربرد بیشـــتری دارد. مراکز اولیه که برای

خو شه‌ها تعیین می شود، بر عملکرد این الگوریتم تاثیر می‌گذارد و براحتی راه‌حل‌های بهینه محلی را ارائه می‌کند. در این الگوریتم زمانی که داده‌ها بزرگ هســـتند رســـیدن به جواب به زمان زیادی نیاز دارد.

۱-۲-۳ طبقه‌بندی

طبقه‌بندی در واقع ارزشیابی ویژگی‌های مجموعه ای از داده‌ها و سپس اختصاص دادن آن‌ها به مجموعه‌ای از گروه‌های از پیش تعریف شده ا ست. م سائل طبقه‌بندی به شنا سایی خ صو صیاتی منجر می شوند که م شخص می‌نمایند هر مورد به کدام گروه تعلق دارد. این الگو هم می‌تواند برای فهم داده موجود و هم برای پیش‌بینی این‌که هر نمونه جدید چگونه کار می کند استفاده شود.

۱-۲-۴ طبقه‌بندی در مقابل خوشه‌بندی

درک تفاوت بین خوشه‌بندی و طبقه‌بندی بسیار مهم می‌باشد. در طبقه‌بندی، همه دسته‌ها و ویژگی‌های آنها از همان ابتدا مشخص است و ب ـرای ه ـر ش ـی تنه ـا کافیست بررسی کنیم که دارای ویژگی‌های کدام طبقه می باشد، اما در خوشه‌بندی هیچ اطلاعاتی در مورد خوشه‌ها نـــداریم و هنگام بررسی اشیاء، بایستی مدلی برای شباهت بین داده‌ها استخراج گردد. در واقع در اغلب الگوریتم‌های خوشه‌بنـدی در ابتدا بایستی راهی برای درک الگوی مابین داده‌ها به دست آورد.

۱-۲-۵ هدف از خوشه‌بندی چیست؟

هدف از خو شه‌بندی یافتن خو شه‌های م شابه از ا شیاء در بین نمونه‌های ورودی می با شد، اما چگونه می توان گفت که یک خو شه‌بندی منا سب ا ست و دیگری منا سب نی ست؟ می توان ن شان داد که هیچ معیار مطلقی برای بهترین خو شه‌بندی وجود ندارد، بلکه این ب ستگی به م ساله و نظر کاربر دارد که باید ت صمیم بگیرد که آیا نمونه‌ها به‌در ستی خو شه‌بندی شده‌اند یا خیر. با این حال معیارهای مختلفی برای خوب بودن یک خوشـــه‌بندی ارائه شـــده اســـت که می تواند کاربر را برای رسیدن به یک خوشه‌بندی مناسب راهنمایی کند که در بخش‌های بعدی چند نمونه از این معیارها

آورده شده است. یکی از مسایل مهم در خوشه‌بندی انتخاب تعداد خوشه‌ها می باشد. در بعضی از الگوریتم‌ها تعداد خوشه‌ها از قبل مشخص شده است و در بعضی دیگر خود الگوریتم تصمیم می گیرد که داده‌ها به چند خوشه تقسیم شوند.

۱-۲-۶ مراحل خوشه‌بندی

مراحل خوشه‌بندی به صورت زیر می باشد:

۱- نمایش الگو که شامل استخراج و انتخاب ویژگی است.

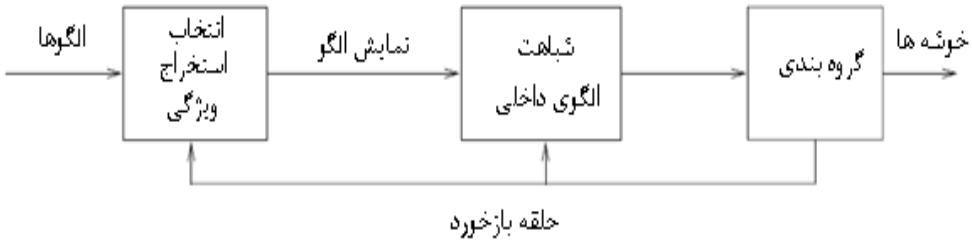
۲- تعریف از اندازه‌گیری شباهت الگوها برای دامنه داده‌ها

۳- خوشه‌بندی یا گروه‌بندی

۴- انتزاع داده (در صورت نیاز)

۵- ارزیابی خروجی (در صورت نیاز)

شکل زیرترتیب سه مرحله اولیه خوشه‌بندی را نشان می‌دهد که شامل بازخوردی از مرحله گروه‌بندی به مراحل قبلی می‌باشد.



شکل ۱-۲: مراحل فرآیند خوشه بندی

۱. نمایش الگو: شامل مشخص کردن تعداد کلاس‌ها و الگوهای در دسترس، انواع و میزان اهمیت ویژگی‌ها برای خوشه‌بندی است.

۲. انتخاب ویژگی: فرآیندی است که تاثیرگذارترین زیر مجموعه‌ها از ویژگی‌های اصلی را برای خوشه‌بندی استفاده می‌نماید.

۳. شباهت الگو: معمولا به وسیله تابع فاصله برای هر زوج الگو، میزان شباهت، اندازه‌گیری می‌شود. معیارهای مختلفی برای اندازه‌گیری فاصله میان الگوها استفاده می‌شود که

معروف‌ترین آن فاصله اقلیدسی است.

۴. خوشه‌بندی: بر اساس روش‌های مختلفی انجام می‌گیرد. مرزبندی میان خوشه‌ها می‌تواند سخت و یا به صورت فازی به شکلی که هر الگو درجه عضویت متفاوتی نسبت به هر گروه داشته باشد، بررسی گردد.

۵. انتزاع داده: فرآیندی است که یک نمایش ساده و فشرده از مجموعه داده‌ها، را استخراج می‌نماید. در واقع در خوشه‌بندی، انتزاع داده توصیف فشرده ای از هر خوشه است.

۶. تحلیل خوشه‌بندی: به دنبال سازمان‌دهی مجموعه‌ای از داده‌ها در یک سری خوشه اتفاق می‌افتد[۲۰].

۱-۲-۷ معیارهای ارزیابی خوشه‌بندی

برای استفاده از خوشه‌بندی در حل مسائل، بایستی آن را در دنیای ریاضی بررسی نمائیم. به همین دلیل در این بخش، نحوه تعریف خوشه‌بندی در ریاضیات مورد توجه قرار گرفته است.

در ریاضیات، یک مساله خوشه‌بندی به صورت ترکیبی از عناصر $X = \{x_1, \ldots, x_n\}$ که نشان دهنده مجموعه ای از n بردار در فضای ویژگی S است، در نظر گرفته می‌شود.

هدف یک مساله خوشه بندی، یافتن یک افراز بهینه $U^* = \{C_1^*, \ldots, C_K^*\}, C_i \cap C_j = \emptyset$ برای مجموعه X است، که C_i^* نمادی برای i امین خوشه از افراز U^* می‌باشد. برای هر مساله تابع هدفی تعریف و $m(u)$ نامیده می‌شود که برحسب آن، تا حد امکان الگوهای مربوط به یک خوشه، یکسان و الگوهای مربوط به خوشه‌های مختلف، متفاوت می‌باشد.

۱-۳-اندازه‌گیری فاصله در مسائل خوشه‌بندی

در خوشه‌بندی، اندازه‌گیری فواصل یکی از مسائل کلیدی است، زیرا شباهت بین دو بردار مختلف x_i و x_j از طریق میزان فاصله‌شان از یکدیکر در فضای ویژگی S سنجیده می‌شود.

اغلب رایج است که برای محاسبه فاصله از نرم اقلیدسی، رابطه زیراستفاده می گردد: