
بسم الله الرحمن الرحيم

دانشگاه تبریز

دانشگاه تبریز

دانشکده‌ی ریاضی

گروه علوم کامپیوتر

پایان نامه

جهت دریافت درجه کارشناسی ارشد در رشته علوم کامپیوتر

عنوان

بررسی الگوریتم های خوشه بندی با رویکرد بهبود در دقت آنها

استاد راهنما:

دکتر محمد رضا فیضی درخشی

استاد مشاور:

میر محمد اتفاق

پژوهشگر:

فاطمه محمودلو

شهریور ۱۳۹۲

نام خانوادگی دانشجو: **محمودلو** نام: **فاطمه**

عنوان پایان‌نامه: **بررسی الگوریتم‌های خوشه‌بندی با رویکرد بهبود در دقت آنها**

استاد راهنما: **دکتر محمدرضا فیضی درخشی**

استاد مشاور: **دکتر میر محمد اتفاق**

مقطع تحصیلی: **کارشناسی ارشد** رشته: **علوم کامپیوتر** گرایش: **سیستم‌های هوشمند**

دانشگاه: **دانشگاه تبریز** دانشکده: **علوم ریاضی**

تاریخ فارغ التحصیلی: **۱۳۹۲/۰۶/۱۶** تعداد صفحه: ۸۲

کلید واژه‌ها: خوشه‌بندی، الگوریتم جنگل و الگوریتم‌های فراابتکاری

چکیده: خوشه‌بندی قرار دادن داده‌ها در گروه‌هایی است، که اعضای هر گروه از زاویه‌ی خاصی شبیه یکدیگرند، بطوریکه شباهت درون هر خوشه حداکثر و شباهت بین داده‌های درون خوشه‌های متفاوت، حداقل باشد خوشه‌بندی فرآیند یادگیری . بدون ناظر است که از قبل هیچ دانشی درباره برچسب داده‌ها ندارد. روش‌های زیادی برای خوشه‌بندی وجود دارد که می‌توان آنها را به روش‌های افرازبندی و سلسله مراتبی تقسیم کرد. در این پایان نامه از الگوریتم فراابتکاری جدیدی به نام الگوریتم جنگل که از طبیعت جنگل الهام گرفته شده، برای خوشه‌بندی استفاده شده است.در این الگوریتم، برای رهایی از بهینه‌های محلی، تغییراتی در قسمت دانه پراکنی محلی انجام شد و نتایج خوبی بدست آمد. برای ارزیابی، روش ارائه شده را روی دو سری مجموعه داده مورد آزمایش قرار دادیم. سری اول شامل داده‌های استاندارد و سری دوم شامل مجموعه داده‌های حقیقی استخراج شده جهت عیب یابی سیستم‌های دوار مکانیکی، است. همچنین از روش‌های شناخته شده‌ای نظیر روش‌های GA، PSO، ACO، CAS_C و K-means برای مقایسه نتایج بدست آمده استفاده شده است. این الگوریتم برای داده‌های استاندارد کمترین مقدار مجموع مجذور فاصله درون خوشه‌ها را بدست آورد. این مقدار در داده iris برابر ۹۶.۶۵۵۷ با انحراف معیار ۰.۰۰۱، در wine برابر ۱۶۲۹۲.۴۱۰۰ با انحراف معیار ۵.۳۴۷۴ و در داده‌های Glass برابر ۲۱۰.۵۳۴۰ با انحراف معیار ۱.۸۰۲۹ است که در مقایسه با دیگر روش‌ها، روش فوق نتیجه مطلوبی را تولید می‌کند. همچنین در این روش فاصله درون خوشه‌ها کاهش و فاصله بیرون خوشه‌ها افزایش پیدا کرده است. علاوه بر این روش فوق درصد خطای خوشه‌بندی پایین‌تری نسبت به دیگر روش‌ها دارد.

ماحصل آموخته‌هایم را تقدیم می‌کنم به آنان که مهر آسمانی‌شان آرام‌بخش آلام زمینی‌ام است

به استوارترین تکیه‌گاه، دستان پرمهر پدرم

و به دلسوزترین نگاه، چشمان شورانگیز مادرم

که هرچه آموخته‌ام در مکتب عشق شماست و هرچه بکوشم قطره‌ای از دریای بی‌کران عشق‌تان را سپاس‌گزار نیستم

باشد که حاصل تلاشم نسیم‌گونه غبار خستگی‌تان را بزداید

و تقدیم به خواهرم

به همسفر مهربان زندگی‌ام، شهربانوی نازنینم

که با هم آغاز کردیم، در کنار هم آموختیم و به امید هم به آینده چشم می‌دوزیم. قلبم لبریز از عشق به توست و خوشبختی‌ات منتهای آرزویم.

فهرست مطالب

عنوان | شماره‌ی صفحه

فهرست اشکال

عنوان | شماره‌ی صفحه

فهرست جداول

عنوان | شماره‌ی صفحه

فصل اول:

مقدمه

۱-۱ مقدمه

خوشه‌بندی[1]، دسته‌بندی بدون ناظر الگوها (مشاهدات، داده‌ها یا بردار ویژگی‌ها) درون گروه‌ها (خوشه‌ها) است [۱]. امروزه با توجه به رشد روز افزون داده‌ها، خوشه‌بندی یکی از ایده‌آل ترین مکانیزم‌ها برای ورود به دنیای عظیم داده‌ها است چرا که تشخیص ساختار داده‌ها را امکان پذیر می‌کند. دلیل اهمیت خوشه‌بندی این است که یک دید کلی و سریع از یک مجموعه بزرگی از اسناد را در اختیار کاربر قرار می‌دهد. می‌توان خوشه‌بندی را به این صورت تعریف کرد که خوشه‌بندی قرار دادن داده‌ها درون خوشه‌ها است بطوریکه داده‌هایی که در یک خوشه قرار دارند بیشترین شباهت را با یکدیگر و با داده‌های خوشه‌های دیگر کمترین شباهت را داشته باشند.

۱-۲ شرح مسئله

خوشه‌بندی روشی شناخته شده برای دریافت مفاهیم پنهان در متن داده‌ها است. بحث اصلی خوشه‌بندی، تقسیم داده‌ها به گروه‌هایی از عناصر است که شبیه یکدیگر باشند. هر کدام از این گروه‌ها که خوشه نامیده می‌شود، شامل عناصری است که با یکدیگر شباهت دارند و با عناصر سایر گروه‌ها متفاوتند.

خوشه‌بندی را می‌توان، یافتن ساختاری در مجموعه‌ای از داده‌ها دانست که دسته‌بندی نشده‌اند. به بیان دیگر قرار دادن داده‌ها در گروه‌هایی است که اعضای هر گروه از زاویه‌ی خاصی شبیه یکدیگرند، بطوریکه شباهت درون هر خوشه حداکثر و شباهت بین داده‌های درون خوشه‌های متفاوت حداقل باشد. برای انجام این کار روش‌های زیادی وجود دارد که می‌توان آنها را به دو دسته روش‌های افرازبندی و سلسله مراتبی تقسیم کرد. امروزه مسئله خوشه‌بندی به یک مسئله بهینه سازی تبدیل شده، که هدف آن پیدا کردن مراکزی است که مجموع فاصله آنها با دیگر داده‌ها حداقل شود. از این‌رو می‌توان از الگوریتم‌های

[1] clustering

فراابتکاری[1] برای حل آن استفاده کرد؛ به عبارتی هدف این الگوریتم‌ها پیدا کردن مراکز خوشه‌ها است بطوری که فاصله آنها از دیگر داده‌ها حداقل شود.

در این پایان نامه برای حل مسئله خوشه‌بندی، از الگوریتم فراابتکاری جدیدی به نام الگوریتم جنگل استفاده شده است. برای بررسی کارایی الگوریتم ارائه شده از دو سری مجموعه داده که اولی مجموعه داده‌های استاندارد و دیگری مجموعه داده‌های حقیقی استخراج شده برای عیب‌یابی سیستم‌های دوار مکانیکی استفاده شده است.

۱-۳ ساختار پایان نامه

در ادامه مبحث، در فصل دوم، مسائل مربوط به خوشه‌بندی و مفاهیم پایه‌ای لازم برای خوشه‌بندی توضیح داده شده است. در ابتدا در مورد مفهوم خوشه‌بندی و اینکه خوشه‌بندی چیست و چه کاربردهایی دارد، توضیح مختصری ارئه شده؛ سپس تفاوت آن با طبقه‌بندی ذکر و اهداف خوشه‌بندی و مراحل انجام آن شرح داده شده است. برای خوشه‌بندی روش‌های زیادی وجود دارد که می‌توان آنها را به دو دسته روش‌های افرازبندی و سلسله مراتبی تقسیم نمود؛ هر یک از این روش‌ها مزایا و معایبی دارند که در ادامه فصل دوم اشاره‌ی مختصری به آنها شده است. . برای انجام خوشه‌بندی الگوریتم‌های زیادی ارائه شده است[۵] که در ادامه این فصل اشاره کوتاهی به برخی از این الگوریتم‌ها خواهیم داشت. بعد از اینکه خوشه‌بندی انجام شد معیارهایی مورد نیاز است تا بتوان خوشه‌بندی انجام شده را ارزیابی کرد، تعدادی از این معیارها در انتهای فصل دوم آورده شده است.

همانطور که ذکر شد یکی از روش‌های خوشه‌بندی، خوشه‌بندی افرازبندی است؛ روش‌هایی که به صورت افرازبندی خوشه‌بندی را انجام می‌دهند خود به دو دسته تقسیم می‌شود، روش‌های کلاسیک و روش‌های فراابتکاری. برای هر دو این روش‌ها الگوریتم‌های بسیاری وجود دارد که در فصل سومri به مطالعه

[1] heu

و بررسی تعدادی از آنها خواهیم پرداخت.

یکی از الگوریتم‌های فرامکاشفه‌ای جدید الگوریتم جنگل می‌باشد؛ این الگوریتم نیز همانند دیگر الگوریتم‌های فرامکاشفه‌ای مبتنی بر جمعیت است. در ابتدای فصل چهارم نگاه اجمالی به الگوریتم جنگل داریم و سپس نحوه بکار گیری این الگوریتم را در مسئله خوشه‌بندی (FAC) مورد بررسی قرار دادیم. داده هایی که برای ارزیابی این الگوریتم استفاده شده، شامل دو سری داده است، که سری اول شامل داده‌های استاندارد و سری دوم شامل داده‌های حقیقی می‌باشد که برای عیب یابی سیستم‌های دوار مکانیکی تولید شده‌اند. در ادامه فصل چهارم این داده‌ها را معرفی کرده‌ایم و نتایج آنها را در جداول جداگانه آوردیم.

در نهایت در فصل پنجم یک نتیجه‌گیری کلی از روش ارائه شده خواهیم داشت و در نهایت راه‌کارهایی را برای کارهای آتی ارائه دادیم.

فصل دوم:

بررسی مفاهیم و الگوهای اولیه

۲-۱ مقدمه

امروزه با حجم عظیمی از داده‌ها روبرو هستیم که برای استفاده از آنها به ابزارهای کشف دانش نیاز داریم. داده‌کاوی به عنوان یک توانایی پیشرفته، در تحلیل داده و کشف دانش مورد استفاده قرار می‌گیرد. برای این کار روش‌های متعددی وجود دارد که هر یک از آنها برای اهدف خاصی مورد استفاده قرار می‌گیرند. یکی از مهمترین روش‌های داده‌کاوی، خوشه‌بندی است که کاربرد بسیاری در کشف دانش دارد.

۲-۲ خوشه‌بندی

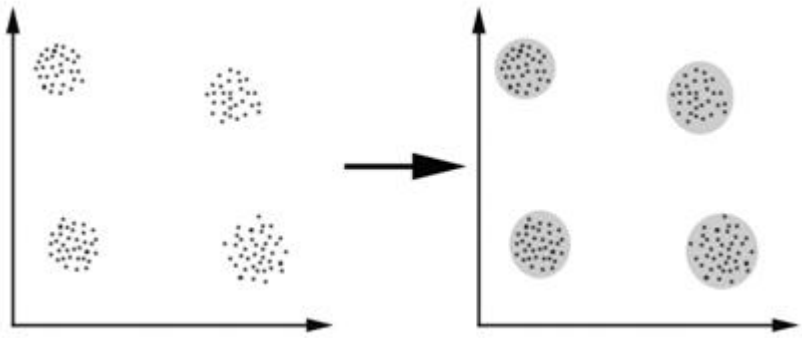
یکی از مراحل پایه‌ای درک و یادگیری داده‌ها، گروه‌بندی آن‌ها در دسته‌های ملموس است. در حالت پایه، سیستم‌های گروه‌بندی داده، اگر اطلاعاتی درباره گروه داده‌ها را در اختیار داشته باشند، با ناظر، در غیر این صورت بدون ناظر نامیده می‌شوند[۶]. خوشه‌بندی یکی از شاخه‌های یادگیری بدون ناظر می‌باشد و فرآیند خودکاری است که نمونه ها را به دسته‌هایی که اعضای آن مشابه یکدیگر هستند، تقسیم می‌کند؛ که به این دسته‌ها خوشه گفته می‌شود.

۲-۲-۱ تعریف خوشه‌بندی

خوشه‌بندی یک تکنیک دسته‌بندی بدون نظارت است که مجموعه‌ی داده‌ها که معمولا بردارهایی در فضای چند بعدی هستند را براساس یک معیار شباهت یا عدم شباهت به تعداد مشخصی خوشه تقسیم می‌کند [۷]؛ بنابراین خوشه مجموعه ای از اشیاء است که در آن اشیاء با یکدیگر مشابه بوده و با اشیاء موجود در خوشه‌های دیگر غیر مشابه هستند. برای مشابه بودن می‌توان معیارهای مختلفی در نظر گرفت؛ مثلا می‌توان معیار فاصله را برای خوشه‌بندی مورد استفاده قرار داد و اشیائی که به یکدیگر نزدیکتر هستند را به عنوان یک خوشه در نظر گرفت که به این نوع خوشه‌بندی، خوشه‌بندی مبتنی بر فاصله[۳] می‌گویند. به عنوان مثال در شکل ۲-۱ نمونه‌های ورودی در سمت چپ به چهار خوشه مشابه شکل سمت راست تقسیم

[3]Distance-based Clustering

می‌شوند. در این مثال هر یک از نمونه‌های ورودی به یکی از خوشه‌ها تعلق دارد و نمونه‌ای وجود ندارد که متعلق به بیش از یک خوشه باشد.



شکل ۲-۱: خوشه‌بندی نمونه‌های اولیه[]

ولی برخی شباهت‌ها را نمی توان به صورت عددی تخمین زد، لذا معیار شباهت دیگر، شباهت مفهومی نام دارد که به این نوع از خوشه‌بندی، خوشه‌بندی مفهومی گفته می‌شود. در خوشه‌بندی مفهومی، خوشه‌ها تنها گروهی از اشیاء با شباهت عددی نیستند، بلکه خوشه‌ها به عنوان گروهی از اشیا که با یکدیگر یک مفهوم را نشان می‌دهند، هستند؛ در واقع تنها تعدادی خوشه تولید می‌شود که مفاهیم مرتبط را توصیف می‌کند. برای خوشه‌بندی مفهومی، ما به مجموعه‌ای از صفات برخی اشیاء (یک زبان توصیف برای مشخص کردن خوشه‌های چنین اشیایی) و یک معیار کیفیت خوشه‌بندی نیاز داریم. هدف، تقسیم‌بندی کردن اشیاء در خوشه‌ها به گونه‌ای است که معیار کیفیت ماکزیمم شود و در عین حال توصیفات عمومی از این خوشه‌ها را تعیین کند. توجه شود که در خوشه‌های مفهومی خصوصیات خوشه، با بررسی و دقت به فرآیند مشخص کردن خوشه‌ها، به وجود می‌آید. این یک تفاوت اصلی بین خوشه‌بندی مفهومی و خوشه‌بندی مبتنی بر فاصله است. در روش خوشه‌بندی مبتنی بر فاصله، خوشه‌ها مطابق با یک معیار شباهت مشخص می‌شوند. این معیار شباهت یک تابع است که فقط خصوصیات اشیاء را مقایسه می‌کند. در مقابل، در خوشه‌بندی مفهومی به شرح یا توصیف هم توجه می‌شود.

۲-۲-۲ کاربردهای خوشه‌بندی

خوشه‌بندی را می‌توان یکی از مهمترین زیر گروه‌های یادگیری بدون نظارت دانست که در موارد بسیاری کاربرد دارد [۸]. به عبارت دیگر خوشه‌بندی، نقش حیاتی در روش‌های طبقه بندی اطلاعات بازی می‌کند برخی از این نمونه‌ها عبارتند از:

- در زمینه مهندسی (مانند مهندسی برق، مهندسی کامپیوتر، یادگیری ماشین، هوش مصنوعی، تشخیص الگو و مهندسی مکانیک) کاربردهایی مانند:
  - داده کاوی[۴]: کشف اطلاعات و ساختار جدید از داده‌های موجود
  - تشخیص گفتار[۵]: در ساخت کتاب کد از بردارهای ویژگی، رد تقسیم کردن گفتار بر حسب گویندگان آن و یا فشرده سازی گفتار
  - تقسیم بندی تصاویر[۶]: تقسیم‌بندی تصاویر پزشکی و یا ماهواره‌ای
  - وب (www): دسته‌بندی اسناد و یا دسته‌بندی مشتریان به سایت‌ها و ...
- علوم پزشکی (شاخه‌های ژنتیک، زیست شناسی، میکروب شناسی، فسیل شناسی، روان شناسی بالینی، آسیب شناسی) مانند:
  - زیست شناسی: دسته‌بندی حیوانات و گیاهان از روی ویژگی‌های آنها
- علوم زمین شناسی (جغرافیا، زمین شناسی، نقشه برداری از زمین) به عنوان نمونه
  - نقشه برداری شهری[۷]: دسته‌بندی خانه‌ها براساس نوع و موقعیت جغرافیایی آنها
  - مطالعات زلزله نگاری[۸]: تشخیص مناطق حادثه خیز براساس مشاهدات قبلی
- علوم اجتماعی (جامعه شناسی، روان شناسی، تاریخ، آموزش و پرورش) که در این راستا به موارد

[4] Data mining
[5] Speech Recognition
[6] Image Segmentation
[7] City-Planning
[8] Earthquake studies

زیر می‌توان اشاره کرد:

- کتابداری: دسته‌بندی کتابها
- بیمه: تشخیص افراد متقلب، تشخیص افرادی که بیمه موتور دارند و بیشترین میزان درخواست از بیمه را نیز در سال مشخصی داشته‌اند.

➢ اقتصاد (بازاریابی، تجارت) به عنوان مثال:

- در بازاریابی[۹]: دسته‌بندی مشتری‌ها به دسته‌هایی بر حسب رفتارها و نیازهای آنها از طریق مجموعه زیادی از ویژگی‌ها و آخرین خریدهای آنها

۲-۲-۳ تفاوت خوشه‌بندی با طبقه‌بندی

برای تقسیم داده در گروه‌های مختلف از دو واژه طبقه‌بندی[۱۰] و خوشه‌بندی استفاده می‌شود. در بسیاری از موارد این دو واژه به جای یکدیگر بکار برده می‌شود. اما به این نقطه باید توجه داشت که اگر چه در هر دو روش به دنبال آن هستیم که داده‌ها را براساس شباهت‌های آنها، در چندین گروه قرار دهیم، ولی طبقه‌بندی با خوشه‌بندی متفاوت است. در طبقه‌بندی نمونه‌های ورودی برچسب گذاری شده‌اند ولی در خوشه‌بندی، نمونه‌های ورودی دارای بر چسب اولیه نمی‌باشند و در واقع با استفاده از روش‌های خوشه‌بندی، داده‌های مشابه، مشخص و بطور ضمنی برچسب‌گذاری می‌شوند. در واقع می‌توان قبل از عملیات طبقه‌بندی داده‌ها، یک خوشه‌بندی روی نمونه‌ها انجام داد و سپس مراکز خوشه‌های حاصل را محاسبه کرد و یک بر چسب به مراکز خوشه‌ها نسبت و سپس عملیات طبقه‌بندی را برای نمونه‌های ورودی جدید انجام داد.

۲-۲-۴ هدف از خوشه‌بندی

هدف روش‌های خوشه‌بندی، گروه‌بندی مجموعه‌ای از داده‌های بدون برچسب است، به طوری که دو داده

[9] Marketing
[10] classification

در یک خوشه تا حد امکان به هم شبیه، و در دو خوشه متفاوت، تا حد امکان از یکدیگر متمایز باشند [۹]. سوالی که مطرح می‌شود این است که چگونه یک خوشه‌بندی مفید داشته باشیم؟ این موضوع به هدف نهایی خوشه‌بندی بستگی دارد. به عبارتی، کاربر باید ملاک را به گونه‌ای تعیین کند که بهترین نتیجه را مطابق با نیازهایش بدست آورد.

خوشه‌بندی برای چهار هدف زیر می‌تواند مورد استفاده قرار بگیرد:

- یافتن ساختار زیربنایی[۱۱] (پیدا کردن شهود راجع به داده‌ها، ایجاد فرضیه، تشخیص ناهنجاری[۱۲] و تعیین ویژگی‌های برجسته)
- گروه‌بندی طبیعی داده‌ها
- فشرده‌سازی به منظور سازمان دهی یا خلاصه‌سازی داده
- کشف گروه‌های طبیعی داده‌ها

۲-۲-۵مراحل انجام خوشه‌بندی

یک روند کلی برای فرایند خوشه‌بندی شامل مراحل زیر است[۶]:

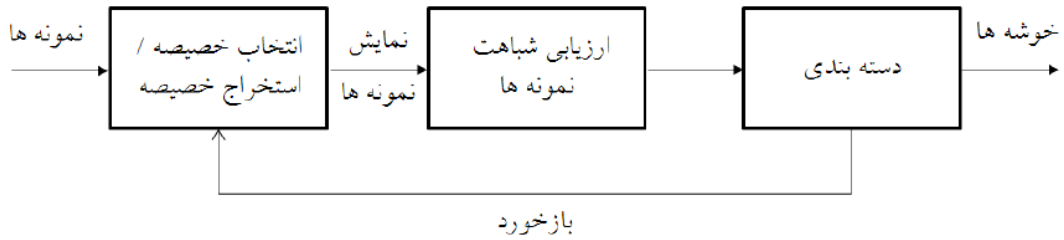
۱. نمایش الگو که معمولا شامل انتخاب یا استخراج خصیصه می‌باشد.
۲. تعریف یک معیار ارزیابی شباهت با توجه به دامنه داده‌ها.
۳. فرایند خوشه‌بندی یا گروه‌بندی.
۴. خلاصه سازی داده‌ها در صورت نیاز.
۵. اعتبارسنجی سیستم.

شکل ۲-۲ سه مرحله از این فرایند، که شامل بازخورد حاصل از نتایج خوشه‌بندی روی دو مرحله

[13] underlying
[14] Anomaly

ابتدایی است، را نشان می‌دهد.

نمایش الگوها،[۱۳] به تعداد کلاس‌ها، تعداد نمونه‌های موجود و تعداد، نوع و مقیاس خصیصه‌های موجود در الگوریتم خوشه‌بندی اشاره می‌کند. برخی از این اطلاعات توسط کاربر کنترل نمی‌باشد.



شکل ۲-۲: مراحل انجام خوشه بندی [۶]

انتخاب خصیصه،[۱۴] فرایند شناسایی یک زیر مجموعه از مؤثرترین خصیصه‌ها برای استفاده در خوشه‌بندی است، و استخراج خصیصه،[۱۵] فرایند تغییر برخی خصیصه‌های موجود و تولید خصیصه‌های جدید می‌باشد. هر دوی این تکنیک ها به منظور دست یافتن به مجموعه مناسبی از خصیصه‌ها و افزایش کارایی خوشه‌بندی است.

مجاورت نمونه‌ها،[۱۶] معمولا بوسیله یک تابع فاصله میان زوج الگوهای ورودی، اندازه‌گیری می‌شود. معیارهای گوناگونی برای اندازه گیری فاصله در حوزه های مختلف استفاده می‌شود [۱۲،۱۱،۱۰]. یک معیار اندازه‌گیری ساده مثل فاصله اقلیدسی معمولا برای نمایش عدم تشابه میان دو الگو بکار می‌رود[۱۳]، همچنین برای تشخیص شباهت‌های مفهومی میان الگوها می‌توان از معیارهای دیگری استفاده کرد.

مرحله گروه‌بندی به روش‌های گوناگونی انجام می‌گیرد. نتایج حاصل از خوشه‌بندی می‌تواند به شکل تقسیم‌بندی سخت (تقسیم داده‌ها درون گروه‌های مجزا) و یا تقسیم‌بندی فازی (هر نمونه با درجه‌های

[13] Pattern Representatio
[14] Feature Selection
[15] Feature Extraction
[16] Pattern Proximity

عضویت متفاوتی در گروه‌های مختلف قرار دارد) باشد. الگوریتم‌های خوشه‌بندی سلسله مراتبی از توابع شباهت برای ترکیب و تقسیم خوشه‌ها استفاده می‌کنند، تا زنجیره‌های تودرتویی از تقسیم‌بندی‌های گوناگون را ارائه دهند. الگوریتم‌های خوشه‌بندی افرازبندی تنها یک تقسیم‌بندی از نمونه‌ها را در اختیار ما می‌گذارند که در ادامه به بررسی هر یک از آنها خواهیم پرداخت.

خلاصه‌سازی داده‌ها،[۱۷] فرایند استخراج یک تعریف مختصر و ساده از مجموعه داده‌های اصلی می‌باشد. در مبحث خوشه‌بندی، نمونه‌ای از خلاصه‌سازی داده را می‌توان یک توصیف کوتاه برای هر خوشه دانست که معمولا آن را مدل پیش الگو[۱۸] می‌نامند، برای مثال مرکز ثقل هر خوشه[۱۱] می‌تواند یک توصیف خلاصه‌ای از نمونه‌های درون آن خوشه باشد.

اعتبار سنجی، نتایج حاصل از الگوریتم‌های خوشه‌بندی را ارزیابی می‌کند. اعتبار سنجی قابل مشاهده بوده[۱۴] و برای تشخیص با معنی بودن خروجی مورد استفاده قرار می‌گیرد. یک ساختار خوشه‌بندی زمانی معتبر است که به صورت شانسی اتفاق نیفتد. زمانیکه از روش‌های آماری برای خوشه‌بندی استفاده می‌شود، اعتبارسنجی نیز با استفاده از روش‌های آماری و آزمایش فرضیات صورت می‌گیرد. دو نوع اعتبارسنجی برای مطالعه وجود دارد. اعتبارسنجی بیرونی،[۱۹] ساختار بدست آمده برای داده ها را با ساختارهای قبل از خوشه‌بندی مقایسه می‌کند. اعتبارسنجی درونی،[۲۰] سعی دارد تا مشخص کند آیا ساختار بدست آمده بطور ذاتی برای داده‌ها مناسب است یا نه.

۲-۳ روش‌های خوشه‌بندی

با توجه به کاربرد و تنوع مسایل، برای خوشه بندی روش‌های مختلفی وجود دارد [۵] که می‌توان آنها را به طور کلی به دو روش عمده دسته‌بندی کرد:

[17] Data Abstraction
[18] Prototype
[19] External Examination
[20] Internal Examination