چکیده

خوشه بندی مقید با استفاده ازیادگیری عمیق

خوشه بندی یکی از مسائل مهم در حوزه یادگیری ماشین می باشد و کاربرد آن در حل مسائل مختلف، روز به روز افزایش پیدا کرده است. از آنجاییکه الگوریتم های خوشه بندی بدون ناظر عمل میکنند برای بهبود آنها سعی شده تا دانش اولیه ای را از داده های خام استخراج کرده و آن دانش را تحت عنوان قید به الگوریتم اعمال کرد. با استفاده از اطلاعات خبرگان یا تخمین توزیع داده ها، می توان قیدهایی را برروی نمونه ها بصورت باید- پیوند و نفی- پیوند استخراج کرد و نتایج تحقیقات حاکی از بهبود قابل توجه دقت خوشه‌بندی‌با استفاده از این روش میباشد. تلاشهای زیادی برروی نحوه اعمال این قیدها در الگوریتم های خوشه بندی انجام شده است ولی الگوریتم های خوشه بندی مقید کنونی پیچیدگی محاسباتی بسیار بالایی دارند. از یک زاویه دیگر، تاثیر بسیار مثبت استخراج/انتخاب/یادگیری ویژگیها در تمامی پژوهشهای وزین خوشه بندی، بصورت تجربی نشان داده است.با توجه به پیشرفت چشمگیر شبکه های عصبی عمیق در تمامی کاربردهای عملی، در این تحقیق هم به منظور یادگیری ویژگیها بصورت غیرخطی و هم به منظور خوشه بندی مقید داده های عملی، یک شبکه عمیق طراحی شده و بصورت بدون سرپرست برای هر دو منظور آموزش دیده است. در این تحقیق، ابتدا از یک شبکه کد کننده و یک شبکه کدگشا بصورت سریال استفاده شده تا ویژگیهای مناسبی را بصورت غیرخطی استخراج کند. سپس ویژگیهای آموزش دیده شده (خرجی کد کننده) به یک شبکه عمیق داده میشود که با کمینه سازی فاصله کولبک لیبر سعی در گنجاندن قیدها به منظور خوشه بندی موثرتری میکند. پس از مرحله آموزش، این روش بسیار سریع در فاز تست عمل میکند و حتی برای داده های با ابعاد بالا نیز از پیچیدگی زمانی مناسبی برخوردار باشد. روش پیشنهادی روی دیتاست های معروف MNIST, USPS و Reuters 10K اعمال شد و نتایج مقایسه ای روش پیشنهادی با روشهای عمیق دیگر حاکی از برتری آماری و معنی دار نتایج حاصله در مقایسه با روشهای رقیب میباشد.

واژگان کلیدی: خوشه بندی مقید،خوشه بندی عمیق، شبکه عصبی عمیق، یادگیری عمیق، خودرمزگذار عمیق

فهرست مطالب

عنوان **صفحه**

فصل اول ۱

مقدمه ۲

۱-۱- تعریف مسئله ۲

۱-۲- چالش های خوشه بندی ۲

۱-۳- چرا یادگیری عمیق برای ما اهمیت دارد؟ ۴

۱-۴- نوآوری ها ۶

۱-۵- ساختار پایان نامه ۷

فصل دوم ۸

مبانی نظری پژوهش ۹

۲-۱- خوشه بندی ۹

۲-۱-۱- خوشه بندی سلسله مراتبی ۱۰

۲-۱-۲- روش های خوشه بندی افرازی ۱۰

۲-۱-۳- روش های خوشه بندی طیفی ۱۱

۲-۱-۴- روش های خوشه بندی مبتنی بر شبکه ۱۱

۲-۱-۵- روش های خوشه بندی مبتنی بر چگالی ۱۲

۲-۲- خوشه بندی مقید ۱۲

۲-۲-۱- روش های مبتنی بر فاصله ۱۵

۲-۲-۲- روش های مبتنی بر قید ۱۵
۲-۲-۳- روش های مبتنی بر ارضای قید ۱۶
۲-۳- مزایا و معایب خوشه بندی مقید ۱۹
۲-۳-۱- مزایای استفاده از قید در خوشه بندی ۲۰
۲-۳-۲- مشکلات استفاده از قید در خوشه بندی ۲۱
۲-۴- شبکه عصبی ۲۲
۲-۴-۱- از نورون تا شبکه عصبی ۲۵
۲-۴-۲- الگوریتم انتشار رو به عقب ۲۶
۲-۵- شبکه عصبی عمیق ۳۰
۲-۵-۱- شبکه های عصبی در هم تنیده ۳۲
۲-۶- مدل های مولد عمیق ۳۴
۲-۶-۱- شبکه های خودرمزگذار ۳۵
فصل سوم ۳۷
کارهای انجام شده ۳۸
۳-۱- شبکه خودرمزگذار رقابتی ۳۸
۳-۱-۱- شبکه مولد رقابتی ۳۸
۳-۱-۲- شبکه خودرمزگذار رقابتی ۴۰
۳-۱-۳- خوشه بندی با شبکه خودرمزگذار رقابتی ۴۱
۳-۲- خوشه بندی یک طیفی مقید ۴۳
فصل چهارم ۴۴
راهکار پیشنهادی ۴۵
۴-۱- مقدمه ۴۵
۴-۲- DEC ۴۶

۴-۲-۱- خودرمزگذار پشته شده ۴۷
۴-۲-۲- Kullback-Leibler divergence (KL) ۴۹
۴-۲-۳- DEC مقدار دهی اولیه پارامترها ۵۰
۴-۲-۴- DEC خوشه بندی با معیار KL ۵۲
۴-۳- تعریف مسئله ۵۴
۴-۴- مدل پیشنهادی ۵۴
فصل پنجم ۵۸
آزمایشات و نتایج ۵۹
۵-۱- مجموعه داده ها ۵۹
۵-۲- پیاده سازی ۶۰
۵-۳- نتایج ۶۱
۵-۴- داده های تولید شده ۶۲
فصل ششم ۷۰
نتیجه‌گیری و پیشنهادها ۷۱
۶-۱- نتیجه گیری ۷۱
۶-۲- کارهای آینده ۷۲
منابع ۷۳

فهرست جدول‌ها

عنوان **صفحه**

جدول ۵-۱- ۶۰
جدول ۵-۲- ۶۲

فهرست شکل‌ها

عنوان **صفحه**

شکل۱-۱- داده ها در مختصات قطبی و دکارتی..... ۵
شکل ۲-۱- قیدهای بایدپیوند و نفی پیوند..... ۱۳
شکل ۲-۲- قید Delta و Epsilon..... ۱۴
شکل ۲-۳- یک نورون عصبی بیولوژیکی..... ۲۳
شکل ۲-۴- مدل نورون مصنوعی..... ۲۴
شکل ۲-۵- چند نمونه از توابع فعال ساز..... ۲۴
شکل ۲-۶- یک شبکه با دو لایه مخفی..... ۲۵
شکل ۲-۷- نحوه تنیده شدن یک هسته با تصویر..... ۳۴
شکل ۳-۱- ساختار شبکه مولد..... ۳۹
شکل ۳-۲- ساختار شبکه خودرمزگذار رقابتی..... ۴۰
شکل ۳-۳- ساختار شبکه خودرمزگذار رقابتی برای خوشه بندی..... ۴۲
شکل ۳-۴- نتایج خوشه بندی داده های MNIST باتعداد خوشه ها ۱۶ ..... ۴۳
شکل ۴-۱- ساختار خودرمزگذار پشته شده..... ۴۸
شکل ۴-۲- خودرمزگذار پشته شده در دولایه..... ۴۹
شکل ۴-۳- ساختار شبکه DEC..... ۵۱
شکل ۴-۴- ساختار شبکه CDEC..... ۵۶
شکل ۵-۱- داده های MNIST قبل و بعد از آموزش..... ۶۳
شکل ۵-۲- داده های Reuters-10k قبل و بعد از آموزش..... ۶۴
شکل ۵-۳- داده های USPS قبل و بعد از آموزش..... ۶۵
شکل ۵-۴- نمودارهای مربوط به accuracy بر روی MNIST..... ۶۶
شکل ۵-۵- نمودارهای مربوط به Loss functions بر روی MNIST..... ۶۷
شکل ۵-۶- نمودارهای مربوط به accuracy بر روی Reuters-10k و USPS..... ۶۸
شکل ۵-۷- نمودارهای مربوط به Loss functions بر روی Reuters-10k و USPS..... ۶۹

فصل اول

مقدمه

۱-۱- تعریف مسئله

یک هدف تحقیق در هوش مصنوعی، خودکار کردن وظایفی می باشد که در حال حاضر نیاز به تخصص انسانی دارد. این اتوماسیون مهم است زیرا موجب صرفه جویی در زمان شده و همچنین مسائلی که قبلا بسیار پیچیده بودند را به مسائل قابل حل تبدیل کرده است. تجزیه و تحلیل داده ها و یا توانایی شناسایی الگوهای معنی دار در داده های حجیم بخش مهمی در حوزه هوش مصنوعی می باشد. الگوریتم های خوشه بندی یک گروه خاص و پر اهمیت از ابزارهای تجزیه و تحلیل داده ها و داده کاوی هستند(K. L. Wagstaff & Cardie, 2002) . کاربرد وسیع خوشه بندی در علوم مختلف، از جمله علوم پزشکی، علوم اجتماعی، اقتصاد و همچنین اهمیت بالای آن در حل مسائل دنیای واقعی باعث شده است که مسئله خوشه بندی همچنان طیف وسیعی از پژوهش ها را به خود اختصاص دهد.

۱-۲- چالش های خوشه بندی

با وجود اینکه مطالعات زیادی در زمینه خوشه بندی صورت گرفته ولی هنوز نیازمند الگوریتم هایی با قابلیتهای جدید می باشد. مسئله خوشه بندی از نگاه های مختلف بررسی شده: اینکه چه خوشه هایی را می توان به عنوان خوشه های مناسب تعیین کرد؟ معیار درست فاصله چیست؟ تاثیر نمونه های گروه در خوشه چگونه است؟ چگونه می توان خوشه ها را ارزیابی کرد؟ و...(Xie, Girshick, & Farhadi, 2016). در واقع یادگیری بدون ناظر اکتشافی و تجربی هست، همچنین جستجوی دانش را از طریق کشف انجام می دهد (K. L. Wagstaff & Cardie, 2002). ارائه یک الگوریتم جامع خوشه بندی که تمام مسائل را حل کند کار دشواری

است، دشوار بودن حل مسئله خوشه بندی را می توان به خاطر ذات بدون ناظر آن و هدفهای متفاوت از خوشه بندی در علوم مختلف دانست. وقتی یادگیری بدون ناظر انجام پذیرد به خاطر عدم آگاهی از مدل واقعی داده ها باعث می شود که خوشه بندی دقت مناسبی را نداشته باشد. از طرفی در مسائل واقعی مشکلاتی از قبیل دردسترس نبودن برچسب داده ها، هزینه بر بودن برچسب زدن به داده ها، زمان بر بودن تعیین و انتساب برچسب، باعث شده است که در فرآیند یادگیری نتوانیم از الگوریتم های با ناظر استفاده کنیم. خوشه بندی مقید راهکاری است که این امکان را فراهم می کند تا اطلاعاتی که در مورد مسئله داریم را بتوانیم در امر یادگیری دخیل کنیم و مطالعات اخیر نشان داده اند که وارد کردن این اطلاعات پیش زمینه در الگوریتم، دقت را تا حد زیادی افزایش می دهد و نقش موثری در ارائه مدل واقعی داده به الگوریتم های خوشه بندی ایفا کند. اطلاعات جانبی به صورت کلی می توانند به صورتهای زیر در الگوریتم های یادگیری ظاهر بشوند:

- برچسب کامل داده ها (دسته بندی)
- برچسب قسمتی از داده ها (دسته بندی نیمه نظارتی)
- قیدهای باید- پیوند و نفی- پیوند همراه با برچسب زوج داده ها (خوشه بندی نیمه نظارتی)
- قیدهای باید- پیوند و نفی- پیوند (خوشه بندی مقید)

مواردی مانند شکل خوشه ها، تعداد اعضای هر خوشه، بیشینه اندازه خوشه ها و تعداد خوشه ها نیز می توانند به عنوان اطلاعات جانبی به الگوریتم های یادگیری ارائه بشوند. مزایای استفاده از اطلاعات جانبی در الگوریتم های یادگیری باعث شده است که توجه بسیاری از پژوهشگران را به خود جلب کند. خوشه بندی مقید از اطلاعات جانبی در سطح نمونه بهره می برد، این اطلاعات به صورت قید های باید- پیوند و نباید- پیوند هستند، هر زوج نمونه که در یک خوشه هستند را باید- پیوند و هر زوجی که نباید در یک خوشه باشند را نباید- پیوند می نامند. این قیدها در هدایت الگوریتم به سمت مدل واقعی داده ها و افزایش دقت الگوریتم بسیار کمک می کنند. خوشه بندی مقید را می توان به عنوان راه حلی برای استفاده موثر از خوشه بندی دانست که نیازمند دانش بیرونی در قالب قیدهای باید- پیوند و نفی- پیوند می باشد(Basu, Davidson, & Wagstaff, 2008) .

تا کنون مطالعات زیادی در زمینه خوشه بندی صورت گرفته است و الگوریتم های مختلفی ارائه گردیده است. خوشه بندی های مرسوم و معروف مانند k-means (MacQueen, 1967) و spectral clustering (Von Luxburg, 2007)، داده ها را با استفاده از ویژگی های خام و بر اساس خصوصیات ذاتی و یا شباهت خوشه بندی می کنند. به هر حال وقتی ابعاد داده زیاد

شود این الگوریتم ها موثر نخواهند بود. انتقال داده ها از فضایی با ابعاد بالا به فضایی با ابعاد پایین به صورتی که خوشه بندی را امکان پذیر کند، یک راه حل بصری هست و به صورت گسترده ای مطالعه شده است. این کار با روشهای کاهش ابعاد مانند PCA (Jolliffe, 2011) امکان پذیر است، اما قابلیت ارائه این مدلهای سطحی محدود می باشد (Guo, Gao, Liu, & Yin, 2017). خوشبختانه یادگیری عمیق پیشرفتهای قابل توجهی در زمینه استخراج ویژگی به دست آورده است و داده ها را از فضایی با ابعاد بالا به فضایی با ابعاد پایین انتقال می دهد به گونه ای که ویژگی های به دست آمده در ابعاد پایین در یادگیری بسیار کارآمد هستند. در هر حال پیشرفت شبکه های عصبی عمیق در حوزه یادگیری با ناظر بوده است اما در سالهای اخیر محققین توانسته اند از این شبکه ها در حوزه یادگیری بدون ناظر نیز بهره ببرند.

در این پایان نامه ما از شبکه های عصبی عمیق استفاده خواهیم کرد و خوشه بندی را با استفاده از این شبکه ها انجام خواهیم داد، ما این نوع از خوشه بندی را خوشه بندی عمیق می نامیم (Guo et al., 2017). و قیدها را در این شبکه وارد خواهیم کرد. و نام شبکه ارائه شده در این پایان نامه را خوشه بندی عمیق مقید می نامیم.

۱-۳- چرا یادگیری عمیق برای ما اهمیت دارد؟

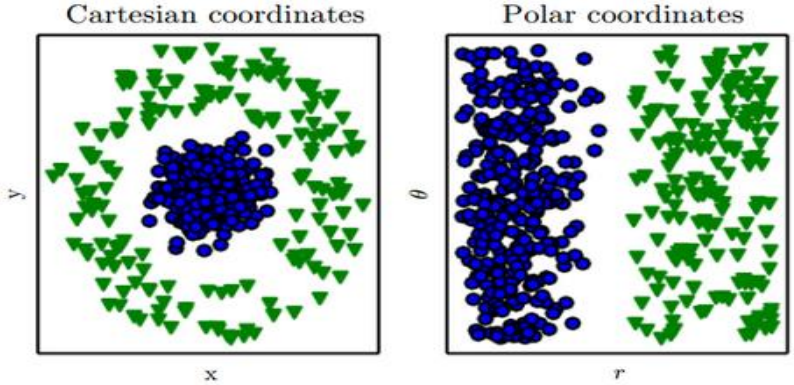
یکی از چالش های خوشه بندی مقید نحوه وارد کردن قید ها به الگوریتم می باشد، به گونه ای که بتواند در ارائه مدل واقعی به الگوریتم کارآمد باشد و فرض های مسئله را بهبود ببخشد و دقت را در خوشه بندی افزایش دهد. مطالعات زیادی در این زمینه صورت گرفته اما تا کنون مطالعات کمی در حوزه یادگیری عمیق انجام شده به گونه ای که مسئله یادگیری و اعمال قیدها به صورت همزمان انجام پذیرد. دلایل اینکه ما تلاش کردیم تا خوشه بندی مقید را با استفاده از شبکه های عمیق انجام دهیم در زیر شرح داده شده است.

چرا یادگیری عمیق برای ما اهمیت دارد؟ میزان کارایی متدهای یادگیری ماشین عمیقا وابسته به فضای ویژگی ای می باشد که در آن به کار برده می شوند. به همین دلیل بیشتر تلاش های واقعی در استقرار الگوریتم های یادگیری ماشین، به سمت طراحی پیش پردازش و انتقال داده به فضایی جدید، پیش رفته تا بتوانند در آن الگوریتم های یادگیری ماشین را پشتیبانی کنند(Bengio, Courville, & Vincent, 2013) و کارایی این الگوریتم ها را بالا ببرند.

چنین مهندسی ویژگی ای برای ما اهمیت دارد ولی کار را سخت میکند و ضعف الگوریتم های جاری را نشان می دهد(نا توانی الگوریتم ها در استخراج داده ها و سازمان دادن اطلاعات متمایز از داده ها.) مهندسی ویژگی یک راه برای به دست آوردن مزایای خلاقیت بشری می باشد و ضعف بیان شده را برطرف می کند. برای گسترش دامنه قابلیت های استفاده از الگوریتم های یادگیری ماشین، اینکه الگوریتم های یادگیری را طوری طراحی کنند که کمتر به منهدسی ویژگی احتیاج نداشته باشند یک ایده آل است و بنابراین برنامه های کاربردی جدید می توانند سریعتر ساخته بشوند، و مهمتر اینکه پیشرفت هایی در هوش مصنوعی ایجاد بشود(Bengio et al., 2013).

در واقع هوش مصنوعی باید اساسا بتواند دنیای اطراف ما را بفهمد، و در (Bengio et al., 2013) استدلال می شود که این تنها زمانی به دست می آید که هوش مصنوعی بتواند یاد بگیرد، تشخیص دهد و عوامل استدلالی پنهان در محیط مشاهده شده از داده های دریافت شده در سطح پایین را کشف کند و در واقع همان استخراج ویژگی را به صورت خودکار انجام دهد.

الگوریتم های یادگیری زمانی کارایی بالایی خواهند داشت که نمایش خوبی از داده به عنوان ورودی در اختیارشان باشد. یادگیری عمیق به صورت سلسله مراتبی و در چندین سطح عمل یادگیری ویژگی را انجام می دهد. از مزایای مهم استخراج ویژگی یا همان بازنمایی داده: اینکه تنک بودن داده منجر به ساده تر شدن بازنمایی داده از خود داده، مقاوم بودن در برابر نویز و تفکیک پذیری بیشتر را موجب می شود. شبکه های عصبی عمیق یکی از قدرتمندترین الگوریتم های یادگیری ماشین در زمینه یادگیری ویژگی به حساب می آیند. در شکل ۱-۱ (Goodfellow, Bengio, Courville, & Bengio,2016) مثالی از تاثیر یادگیری ویژگی نمایش داده شده است.



شکل ۱-۱- (Goodfellow, Bengio, Courville, & Bengio,2016) سمت راست داده ها با استفاده از مختصات قطبی نمایش داده شده اند و شکل سمت چپ داده ها با استفاده از مختصات دکارتی نمایش داده شده اند. در فضای مختصات قطبی داده ها به صورت خطی جداپذیر هستند ولی در شکل سمت چپ جداپذیری به صورت خطی امکان پذیر نیست.

خوشبختانه شبکه های عصبی عمیق گامی بزرگ در این راستا برداشته اند و انقلابی را در یادگیری ویزگی به صورت خودکار برداشته اند. شبکه های عصبی عمیق که از ساختار نورون های مغز الهام گرفته شده اند به خاطر دستاوردهایی که در حوزه یادگیری ماشین و همچنین برنامه های کاربردی به دست آورده اند بسیار مورد توجه قرار گرفته اند. جوهره ی یادگیری ویژگی، خودرمزگذار ها هستند. چارچوب خودرمزگذار یک خط لوله ی محاسباتی فرام می کند و نه تنها یک توصیف آماری. ما نیز در این پایان نامه از این روش و مزایای آن بهره می بریم و از آن در جهت بهبود روش خوشه بندی مقید استفاده خواهیم کرد.

۱-۴- نوآوری ها

نوآوری ها در این پایان نامه را می توان به صورت زیر عنوان کرد:

- **استفاده از شبکه عصبی عمیق در حل خوشه بندی مقید:** خود مسئله خوشه بندی چون بدون ناظر است مسائلی را به وجود می آورد از جمله اینکه ساختار شبکه های عصبی عمیق از ابتدا به صورت با ناظر طراحی شده است و در اصل یک الگوریتم با نظارت بوده. تعریف تابع هدفی که بتواند مسئله خوشه بندی را با استفاده از شبکه های عمیق انجام بدهد خود چالش بزرگی است. (Xie et al., 2016) مدلی ارائه دادند که در آن شبکه عمیق با استفاده از یک تابع هدف خود آموز و به صورت تکرار شونده، آموزش می بیند و داده ها را خوشه بندی می کند. ما تابع هدفی تعریف می کنیم که قید های باید- پیوند را در این شبکه وارد کند و نشان خواهیم داد که دقت الگوریتم تا حد خوبی بالا می رود. تا زمان ویرایش این پایان نامه هنوز الگوریتمی ارائه نشده که خوشه بندی مقید را به صورت عمیق انجام دهد.
- **ارائه یک تابع هدف جدید:** در این پایان نامه یک تابع هدف تعریف می گردد که می تواند قیدهای باید- پیوند را در یادگیری شبکه عمیق دخیل کند. ما از معیار کولبک لیبلر[1] و انتشار رو به عقب استفاده خواهیم کرد. فرض ما بر این است که قید های باید- پیوند در خروجی شبکه باید توزیعی شبیه به هم داشته باشند و بنابراین برای قیدهای باید- پیوندی که در خروجی شبکه توزیع یکسانی ندارند، خطایی را در شبکه پس انتشار می دهیم. این امر باعث می شود این نمونه ها کنار هم بمانند و فرض های اولیه خوشه بندی را به مدل واقعی داده نزدیکتر کنند.

1 Kullback-Leibler divergence (KL-divergence)

۱-۵- ساختار پایان نامه

در فصل دوم (مبانی نظری پژوهش) خوشه بندی، خوشه بندی مقید و یادگیری عمیق بررسی میشوند، سپس در فصل سوم کارهای انجام شده به طور خلاصه بیان می شود. فصل چهارم راهکار پیشنهادی شرح داده خواهد شد و فصل پنم و ششم آزمایشات و نتایج را بررسی خواهیم کرد.

فصل دوم

مبانی نظری پژوهش

۲-۱- خوشه بندی

خوشه بندی تخصیص دادن مجموعه ای از مشاهدات به زیر گروه ها (خوشه ها) می باشد به صورتی که مشاهدات در یک گروه از نظر بعضی خصوصیات شبیه هم باشند و مشاهدات در گروه های مختلف از نظر همان خصوصیات متفاوت باشند. سه هدف کلی خوشه بندی، کشف ساختار پنهان داده ها ، کاهش ابعاد و پیش پردازش داده ها می باشد. در حالی که خوشه بندی تاریخچه طولانی دارد و تکنیک های زیادی در حوزه های آمار، شناسایی آماری الگو، داده کاوی و دیگر حوزه ها ارائه گردیده است، همچنان چالش های مهمی باقی است (Steinbach, Ertöz, & Kumar, 2004). به طور کلی یافتن راه حل بهینه برای یک مسئله خوشه بندی NP-hard میباشد (Garey & Johnson, 1979). ابهام در تعریف خوشه مناسب، بدون ناظر بودن یادگیری در الگوریتم، مشکل بودن تعریف معیار فاصله از مهمترین دلایل سخت بودن مسئله خوشه بندی می باشد. از روشهای مطرح خوشه بندی می توان به خوشه بندی سلسله مراتبی[۱]، افرازی[۲]، طیفی[۳]، مبتنی بر شبکه[۴] و مبتنی بر چگالی[۵] اشاره نمود (Elavarasi, Akilandeswari, & Sathiyabhama, 2011). چرا خوشه بندی؟ خوشه بندی در تمام رشته هایی که دارای متغیرهای چند مقداری[۶] هستند متداول است. یک جستجو با (Scholar, 2009) نشان می دهد که ۱۶۶۰ ورودی با کلمه خوشه بندی داده ها فقط در سال ۲۰۰۷ ظاهر شده است. این ادبیات گسترده به اهمیت خوشه بندی در تحلیل داده ها اشاره دارد. کار سختی است که به صورت جامع بتوان زمینه های علمی متعدد و برنامه های کاربردی که از تکنیک خوشه بندی استفاده کرده اند و همچنین هزاران الگوریتم منتشر شده را لیست

1 Hierarchical
2 Partitional
3 Spectral
4 Grid based
5 Density based
6 Multivariate