



وزارت علوم، تحقیقات و فناوری

دانشگاه علوم و فنون مازندران

پایان نامه

مقطع کارشناسی ارشد

رشته : فناوری اطلاعات - مدیریت سیستم‌های اطلاعاتی

عنوان : خوشه‌بندی مبتنی بر انتخاب بر اساس نظریه خرد جمعی

استاد راهنما : جناب آقای دکتر بهروز مینایی

استاد مشاور : جناب آقای دکتر حسین علیزاده

دانشجو : محمد یوسف‌نژاد

تابستان ۱۳۹۲

سپاس و آفرین آن پادشا را

که گیتی را پدید آورد و مارا

بسم الله الرحمن الرحيم

خدایا بزرگ و توانا تویی

رحیم و رئوفی و یکتا تویی

پر از مهری و بخشش و مغفرت

که ما قطره هستیم و دریا تویی

هر طرف جویم تو را چون عیان در جان تویی
زان هدف پویم تو را عاشق جانان تویی
تقدیم به

تمام کسانی که با دعای خیرشان مرا یاری کردند،
که این قدردانی کوچکی است که می‌توانم در قبال لطف بی‌دریغشان تقدیم نمایم.

مخصوصاً پدر عزیزم

که همیشه پشتیبان و مشوق من بوده است.

و روح پاک مادر فداکارم

که به من درستکاری و پاک دامنی آموخت.

و همسر مهربانم

که همواره مرا در این راه یاری نموده است.

باقی بقايتان، جانم فدايتان

با سپاس گذاری از

زحمات تمامی اساتید عزیز و بزرگوارم،

مخصوصاً جناب آقای دکتر مینایی و جناب آقای دکتر علیزاده

که با لطف فرآوان و راهنمایی دلسوزانه و تجربیات گرانقدرشان، مرا در

ارایه این تحقیق یاری فرموده اند.

چکیده

خوشبندی وظیفه کاوش الگوهای پنهان در داده‌های بدون برچسب را بر عهده دارد. به خاطر پیچیدگی مسئله و ضعف روش‌های خوشبندی پایه، امروزه روش‌های خوشبندی ترکیبی مورد استفاده قرار می‌گیرند. به روشنی از خوشبندی ترکیبی که در آن از زیرمجموعه‌ای منتخب از نتایج اولیه برای ترکیب و ساخت نتیجه نهایی استفاده می‌شود خوشبندی ترکیبی مبتنی بر انتخاب زیرمجموعه نتایج اولیه می‌گویند. در سال‌های اخیر تمرکز بر روی ارزیابی نتایج اولیه برای انتخاب خوشبندی ترکیبی موردنظر محققین زیادی قرار گرفته است. اما پاسخ به بعضی از سوالات در این زمینه همچنان با ابهامات زیادی روبروست. از طرفی دیگر، نظریه خرد جمعی که اولین بار توسط سورویکی منتشر شده است، نشان می‌دهد که قضاوت‌های جمعی و دموکراتیک از اعتبار بیشتری نسبت به آنچه که ما انتظار داشتیم برخوردار هستند. این نظریه چهار شرط پراکندگی، استقلال، عدم تمرکز و روش ترکیب مناسب آراء را برای هر جمعیت خردمند لازم و کافی می‌داند. هدف این تحقیق پیشنهاد فرآیندی جهت نگاشت و به کارگیری نظریه خرد جمعی در انتخاب زیرمجموعه مناسب در خوشبندی ترکیبی مبتنی بر انتخاب می‌باشد. از این روی در این تحقیق ابتدا با استفاده از تعاریف مطرح شده در نظریه خرد جمعی باز تعریفی متناسب با خوشبندی ترکیبی مبتنی بر انتخاب ارائه می‌شود و بر اساس آن دو روش برای ترکیب این دو مفهوم پیشنهاد می‌شود. در روش پیشنهادی اول الگوریتم‌های خوشبندی اولیه غیر هم نام کاملاً مستقل فرض خواهد شد و برای ارزیابی استقلال الگوریتم‌های هم نام نیاز به آستانه‌گیری می‌باشد. در روش دوم، سعی شده است تا دو بخش از روش اول بهبود یابد. از این روی جهت مدل‌سازی الگوریتم‌ها و ارزیابی استقلال آن‌ها نسبت به هم یک روش مبتنی بر گراف کد الگوریتم ارائه می‌شود و میزان استقلال به دست آمده در این روش به عنوان وزنی برای ارزیابی پراکندگی در تشکیل جواب نهایی مورد استفاده قرار می‌گیرد. جهت بررسی ادعاهای این تحقیق در بخش ارزیابی دقیقت و اطلاعات متقابل نرمال شده‌ی روش‌های پیشنهادی بر روی داده‌های استاندارد با روش‌های پایه، روش ترکیب کامل و چند روش معروف خوشبندی ترکیبی مبتنی بر انتخاب مقایسه می‌شوند که این مقایسه کاراری ببالای روش‌های پیشنهادی این تحقیق در اکثر موارد نسبت به سایر روش‌های مطرح شده را نشان می‌دهد. همچنین در بخش نتیجه‌گیری چندین روش توسعه جهت کارهای آتی پیشنهاد می‌شود.

واژه‌های کلیدی: خوشبندی ترکیبی، خرد جمعی، استقلال الگوریتم‌های خوشبندی، پراکندگی نتایج خوشبندی اولیه، عدم تمرکز در چهارچوب خوشبندی ترکیبی

فهرست مطالب

فصل اول

۱.	مقدمه	۱
۲		۲
۱-۱.	خوشبندی	۱
۲-۱	خوشبندی ترکیبی	۴
۳-۱	خرد جمعی	۴
۴-۱	خوشبندی مبتنی بر انتخاب بر اساس نظریه خرد جمعی	۵
۴-۱-۱	- فرضیات تحقیق	۷

فصل دوم

۲.	مروری بر ادبیات تحقیق	۹
۹		۹
۱-۲	. مقدمه	۹
۲-۲	. خوشبندی	۹
۲-۲-۱	. الگوریتم های خوشبندی پایه	۹
۲-۲-۱-۱	. الگوریتم های سلسله مراتبی	۱۰
۲-۲-۱-۱-۱	. تعاریف و نمادها	۱۱
۲-۲-۱-۱-۲	. الگوریتم پیوندی منفرد	۱۳
۲-۲-۱-۱-۳	. الگوریتم پیوندی کامل	۱۳
۲-۲-۱-۱-۴	. الگوریتم پیوندی میانگین	۱۴
۲-۲-۱-۱-۵	. الگوریتم پیوندی بخشی	۱۵
۲-۲-۱-۲	. الگوریتم های افزایشی	۱۵
۲-۲-۱-۲-۱	. الگوریتم K-means	۱۶
۲-۲-۱-۲-۲	. الگوریتم FCM	۱۷
۲-۲-۱-۲-۳	. الگوریتم طیفی	۱۹
۲-۲-۱-۲-۴	. الگوریتم برش نرمال	۲۰
۲-۲-۱-۲-۵	. الگوریتم NJW	۲۱
۲-۲-۱-۲-۶	. الگوریتم خوشبندی کاہشی	۲۲
۲-۲-۱-۲-۷	. الگوریتم خوشبندی Median K-Flat	۲۳
۲-۲-۱-۲-۸	. الگوریتم خوشبندی مخلوط گوسی	۲۵
۲-۲-۲	. معیارهای ارزیابی	۲۷
۲-۲-۲-۱	. معیار SSE	۲۸
۲-۲-۲-۲	. معیار اطلاعات متقابل نرمال شده	۳۰
۲-۲-۲-۳	. معیار APMM	۳۲

۳-۲. خوشبندی ترکیبی	۳۳
۱-۳-۲. ایجاد تنوع در خوشبندی ترکیبی	۳۴
۳-۲-۱. استفاده از الگوریتم‌های مختلف خوشبندی ترکیبی	۳۵
۲-۱-۳-۲. تغییر پارامترهای اولیه خوشبندی ترکیبی	۳۵
۳-۲-۱-۳-۲. انتخاب یا تولید ویژگی‌های جدید	۳۶
۴-۱-۳-۲. انتخاب زیرمجموعه‌ای از مجموعه داده اصلی	۳۶
۲-۳-۲. ترکیب نتایج با تابع توافقی	۳۷
۳-۲-۱. روش مبتنی بر مدل مخلوط	۳۷
۲-۲-۳-۲. روش مبتنی بر ابر گراف	۴۴
۲-۲-۳-۲. روش CSPA	۴۶
۲-۲-۳-۲. روش HGPA	۴۷
۲-۲-۳-۲. روش MCLA	۴۸
۳-۲-۳-۲. روش‌های مبتنی بر ماتریس همیستگی	۵۰
۱-۳-۲-۳-۲. الگوریتم‌های سلسله مراتبی تراکمی	۵۱
۲-۳-۲-۳-۲. الگوریتم افزاینده گراف با تکرار	۵۲
۳-۲-۳. الگوریتم‌های خوشبندی ترکیبی کامل	۵۶
۲-۳. خوشبندی ترکیبی مبتنی بر انتخاب	۵۷
۴-۲-۱. خوشبندی ترکیبی مبتنی بر انتخاب فرن ولین	۵۷
۴-۲-۱. تعریف معیار کیفیت در روش فرن ولین	۵۷
۴-۲-۱-۲. تعریف معیار پراکندگی در روش فرن ولین	۵۸
۴-۲-۱-۳. راهکار انتخاب خوش برای تشکیل نتیجه نهایی در روش فرن ولین	۵۸
۴-۲-۲. الگوریتم هوشمند طبقه‌بندی مجموعه داده‌ها	۶۰
۴-۲-۳. خوشبندی ترکیبی طیفی مبتنی بر انتخاب بر اساس شباهت	۶۱
۴-۲-۴-۱. معیار ارزیابی Sim در روش پیشنهادی ژیا	۶۱
۴-۲-۴-۲. انتخاب خوشبندی بر اساس قانون نزدیک‌ترین همسایه در روش ژیا	۶۲
۴-۲-۴-۳. خوشبندی ترکیبی انتخابی لی مین	۶۴
۴-۴-۱. انتخاب افزای مرجع در روش لی مین	۶۴
۴-۴-۲. راهکار انتخاب خوش برای مین در روش لی مین	۶۶
۴-۴-۳. چهارچوب الگوریتم خوشبندی انتخابی لی مین	۶۸
۴-۴-۵. خوشبندی بر اساس معیار MAX با استفاده از مجموعه‌ای از خوش‌های یک افزای	۷۹
۴-۵-۱. راهکار ارزیابی خوشبندی MAX	۷۹
۴-۵-۲. روش انباست مدارک توسعه‌یافته	۷۰
۴-۶. خوشبندی بر اساس معیار APMM با استفاده از مجموعه‌ای از خوش‌های یک افزای	۷۰
۵-۲. روش بهترین افزای توافقی اعتبارسنجی شده	۷۲
۶-۲. استفاده از نظریه خرد جمعی در علوم رایانه	۷۳

فصل سوم

۷۶.....	۳. روش تحقیق.....
۷۶.....	۱-۳. مقدمه.....
۷۷.....	۲-۳. نظریه خرد جمعی.....
۷۸.....	۱-۲-۳. شرایط جامعه خردمند.....
۷۸.....	۱-۲-۳. ۱. تعریف معیار پراکندگی.....
۷۹.....	۱-۲-۳. ۲. تعریف معیار استقلال.....
۷۹.....	۱-۲-۳. ۳. تعریف معیار عدم تمکن.....
۸۰.....	۴-۳. روش ترکیب مناسب.....
۸۰.....	۲-۳. اهمیت و رابطه استقلال و پراکندگی در خرد جمعی.....
۸۲.....	۳-۳. استثناءها در خرد جمعی.....
۸۲.....	۳-۳. خوشبندی خردمند با استفاده از آستانه‌گیری.....
۸۴.....	۱-۳-۳. ۱. روش ارزیابی پراکندگی نتایج.....
۸۵.....	۱-۳-۳. ۲. روش ارزیابی استقلال الگوریتمها.....
۸۶.....	۱-۳-۳. ۳. عدم تمکن در بخش‌های سازنده خوشبندی ترکیبی.....
۹۰.....	۴-۳-۳. مکانیزم ترکیب مناسب.....
۹۰.....	۴-۳-۳. ۵. بررسی تأثیر مکانیزم بازخورد در کیفیت نتیجه نهایی.....
۹۱.....	۴-۳-۳-۶. شبیه کد خوشبندی خردمند با استفاده از آستانه‌گیری.....
۹۳.....	۴-۳. خوشبندی خردمند مبتنی بر گراف استقلال الگوریتم.....
۹۳.....	۴-۳-۱. بررسی مکانیزم حل مسائل توسط الگوریتم‌های خوشبندی.....
۹۵.....	۴-۳-۲. مدل‌سازی گراف استقلال الگوریتم.....
۹۶.....	۴-۳-۲-۴-۳. ۱. زبان استقلال الگوریتم خوشبندی.....
۹۹.....	۴-۳-۲-۴-۳. ۲. تبدیل کد به گراف استقلال الگوریتم.....
۱۰۷.....	۴-۳-۲-۴-۳. ۳. ارزیابی گراف استقلال الگوریتم.....
۱۱۰.....	۴-۳-۴-۳. ۳. چهارچوب خوشبندی خردمند مبتنی بر گراف استقلال الگوریتم.....
۱۱۰.....	۴-۳-۴-۳. ۱. ارزیابی استقلال الگوریتم.....
۱۱۲.....	۴-۳-۴-۳. ۲. روش انباشت مدارک وزن‌دار.....
۱۱۲.....	۴-۳-۴-۳. ۳. شبیه کد خوشبندی خردمند مبتنی بر گراف استقلال الگوریتم.....

فصل چهارم

۱۱۶.....	۴. پیاده‌سازی و تحلیل نتایج.....
۱۱۶.....	۱-۴. مقدمه.....
۱۱۶.....	۲-۴. مجموعه داده.....

۱۱۸	۳-۴. مدل‌سازی الگوریتم‌ها به زبان استقلال الگوریتم
۱۲۸	۴-۴. ابزار تحلیلگر کد استقلال الگوریتم
۱۳۰	۴-۵. نتایج آزمایش‌ها
فصل پنجم	
۱۴۰	۵. جمع‌بندی و کارهای آینده
۱۴۰	۱-۵. جمع‌بندی.....
۱۴۱	۲-۵. کارهای آینده.....
۱۴۲	منابع و مأخذ.....

فهرست جداول

فصل سوم

جدول ۳-۱. نگاشت لغات لاتین در خوشه‌بندی ترکیبی به نظریه خرد جمعی ۹۳

جدول ۳-۲. یک نمونه از جدول نگاشت استاندارد کد ۹۸

فصل چهارم

جدول ۴-۱. مجموعه داده ۱۱۷

جدول ۴-۲. لیست مجموعه الگوریتم‌های پایه ۱۱۹

جدول ۴-۳. جدول نگاشت استاندارد کد ۱۲۰

جدول ۴-۴. دقت نتایج این الگوریتم‌های خوشه‌بندی را نسبت به کلاس‌های واقعی داده ۱۳۰

جدول ۴-۵. جدول مقایسه معیار اطلاعات متقابل نرمال شده (NMI) نتایج آزمایش ۱۳۲

فهرست تصاویر و نمودار

فصل دوم

شکل ۲-۱. یک خوشبندی سلسله مراتبی و درخت متناظر ۱۰
شکل ۲-۲. ماتریس مجاورت ۱۱
شکل ۲-۳. رابطه دودویی و گراف آستانه ۱۲
شکل ۲-۴. گراف‌های آستانه برای ماتریس D_1 ۱۲
شکل ۲-۵. الگوریتم خوشبندی سلسله مراتبی تراکمی پیوندی منفرد ۱۳
شکل ۲-۶. دندوگرام پیوندی منفرد برای ماتریس D_1 ۱۳
شکل ۲-۷. الگوریتم خوشبندی سلسله مراتبی تراکمی پیوندی کامل ۱۴
شکل ۲-۸. دندوگرام پیوندی کامل برای ماتریس D_1 ۱۴
شکل ۲-۹. الگوریتم خوشبندی افزاینده $K-means$ ۱۶
شکل ۲-۱۰. الگوریتم فازی خوشبندی FCM ۱۸
شکل ۲-۱۱. خوشبندی کاهشی ۲۳
شکل ۲-۱۲. شبکه کد الگوریتم MKF ۲۶
شکل ۲-۱۳. (الف) مجموعه داده با تعداد ۱۰ خوش واقعی. (ب) منحنی SSE ۲۹
شکل ۲-۱۴. (الف) مجموعه داده (ب) منحنی SSE مربوطه ۲۹
شکل ۲-۱۵. دو افزای اولیه با تعداد سه خوش ۳۱
شکل ۲-۱۶. نمونه‌های اولیه در نتایج الگوریتم $K-means$ ۳۶
شکل ۲-۱۷. زیر شبکه کد الگوریتم خوشبندی ترکیبی توسط مدل مخلوط ۴۳
شکل ۲-۱۸. خوشبندی ترکیبی ۴۴
شکل ۲-۱۹. نمونه ماتریس $H^{(q)}$, جهت تبدیل خوشبندی به ابر گراف ۴۵
شکل ۲-۲۰. ماتریس شباهت بر اساس خوش برای مثال شکل (۵-۳) ۴۶

۴۷	شکل ۲-۱. الگوریتم افزایش‌بندی ابر گراف
۴۹	شکل ۲-۲. الگوریتم فرا خوش‌بندی
۵۰	شکل ۲-۳. الگوریتم خوش‌بندی ترکیبی مبتنی بر ماتریس همبستگی
۵۳	شکل ۲-۴. الگوریتم افزایش‌بندی با تکرار
۵۴	شکل ۲-۵. نمایش گراف مجاورت در مراحل کاهش درجه ماتریس و شمارش آن
۵۵	شکل ۲-۶. مثال روند تغییر توزیع تعداد خوش
۵۵	شکل ۲-۷. جریان کار عمومی برای پیاده‌سازی الگوریتم افزایش‌بندی گراف
۶۲	شکل ۲-۸. گراف تابع <i>Sim</i> در بازه بین صفر و یک
۶۳	شکل ۲-۹. الگوریتم خوش‌بندی ترکیبی طیفی مبتنی بر انتخاب بر اساس شباهت
۶۶	شکل ۲-۱۰. مثالی از ماتریس اتصال
۶۸	شکل ۲-۱۱. شبکه کد خوش‌بندی ترکیبی انتخابی لی مین
۶۹	شکل ۲-۱۲. روش ارزیابی خوشی یک افزار در روش MAX
۷۱	شکل ۲-۱۳. چهارچوب خوش‌بندی ترکیبی مبتنی بر انتخاب با استفاده از مجموعه‌ای از خوش‌های یک افزار
۷۲	شکل ۲-۱۴. چهارچوب روش بهترین افزار توافقی اعتبارسنجی شده
۷۲	فصل سوم
۸۲	شکل ۳-۱. چهارچوب الگوریتم خوش‌بندی خردمند با استفاده از آستانه‌گیری
۸۶	شکل ۳-۲. محاسبه درجه استقلال دو خوش‌بندی
۸۹	شکل ۳-۳. تأثیر عدم تمرکز بر روی پیچیدگی داده
۹۱	شکل ۳-۴. تأثیر انتخاب افزارها در خوش‌بندی ترکیبی مبتنی بر انتخاب بر مقدار NMI ارزیابی شده
۹۲	شکل ۳-۵. شبکه کد خوش‌بندی خردمند با استفاده از آستانه‌گیری
۹۴	شکل ۳-۶. دسته‌بندی الگوریتم‌های خوش‌بندی
۹۸	شکل ۳-۷. کد الگوریتم K-means به زبان استقلال الگوریتم خوش‌بندی
۱۰۰	شکل ۳-۸. تبدیل کدهای شروع و پایان به گراف

..... ۱۰۰	شکل ۳-۸. تبدیل عملگر شرط ساده به گراف
..... ۱۰۱ شکل ۳-۹. تبدیل عملگر شرط کامل به گراف
..... ۱۰۱ شکل ۳-۱۰. تبدیل عملگر شرط تو در تو به گراف
..... ۱۰۲ شکل ۳-۱۱. تبدیل عملگر حلقه ساده به گراف
..... ۱۰۲ شکل ۳-۱۲. تبدیل عملگر حلقه با پوش به گراف
..... ۱۰۳ شکل ۳-۱۳. پیادهسازی شرط ساده بدون هیچ کد اضافی
..... ۱۰۳ شکل ۳-۱۴. پیادهسازی شرط ساده با کدهای قبل و بعد آن
..... ۱۰۴ شکل ۳-۱۵. پیادهسازی شرط کامل
..... ۱۰۴ شکل ۳-۱۶. پیادهسازی شرط تو در تو
..... ۱۰۵ شکل ۳-۱۷. پیادهسازی یک شرط کامل در یک شرط ساده
..... ۱۰۵ شکل ۳-۱۸. پیادهسازی یک شرط کامل در یک شرط کامل دیگر
..... ۱۰۶ شکل ۳-۱۹. پیادهسازی حلقه ساده
..... ۱۰۶ شکل ۳-۲۰. پیادهسازی یک حلقه ساده داخل حلقه‌ای دیگر
..... ۱۰۶ شکل ۳-۲۱. پیادهسازی یک حلقه داخل یک شرط کامل
..... ۱۰۷ شکل ۳-۲۲. پیادهسازی یک شرط کامل داخل یک حلقه ساده
..... ۱۰۸ شکل ۳-۲۳. ماتریس درجه وابستگی کد
..... ۱۰۸ شکل ۳-۲۴. شبکه کد مقایسه محتوای دو خانه از آرایه‌های استقلال الگوریتم
..... ۱۱۰ شکل ۳-۲۵. چهارچوب خوشه‌بندی خردمند مبتنی بر گراف استقلال الگوریتم
..... ۱۱۳ شکل ۳-۲۶. شبکه کد خوشه‌بندی خردمند مبتنی بر گراف استقلال الگوریتم

فصل چهارم

..... ۱۱۸ شکل ۴-۱. مجموعه داده Halfring
..... ۱۲۱ شکل ۴-۲. الگوریتم K-means
..... ۱۲۱ شکل ۴-۳. الگوریتم FCM

..... شکل ۴-۴. الگوریتم Median K-Flats	۱۲۲
..... شکل ۴-۵. الگوریتم Gaussian Mixture	۱۲۲
..... شکل ۴-۶. الگوریتم خوشبندی Subtractive	۱۲۲
..... شکل ۴-۷. الگوریتم پیوندی منفرد با استفاده از معیار فاصله اقلیدسی	۱۲۳
..... شکل ۴-۸. الگوریتم پیوندی منفرد با استفاده از معیار فاصله Hamming	۱۲۳
..... شکل ۴-۹. الگوریتم پیوندی منفرد با استفاده از معیار فاصله Cosine	۱۲۳
..... شکل ۴-۱۰. الگوریتم پیوندی کامل با استفاده از معیار فاصله اقلیدسی	۱۲۴
..... شکل ۴-۱۱. الگوریتم پیوندی کامل با استفاده از معیار فاصله Hamming	۱۲۴
..... شکل ۴-۱۲. الگوریتم پیوندی کامل با استفاده از معیار فاصله Cosine	۱۲۴
..... شکل ۴-۱۳. الگوریتم پیوندی میانگین با استفاده از معیار فاصله اقلیدسی	۱۲۴
..... شکل ۴-۱۴. الگوریتم پیوندی میانگین با استفاده از معیار فاصله Hamming	۱۲۵
..... شکل ۴-۱۵. الگوریتم پیوندی میانگین با استفاده از معیار فاصله Cosine	۱۲۵
..... شکل ۴-۱۶. الگوریتم پیوندی بخشی با استفاده از معیار فاصله اقلیدسی	۱۲۵
..... شکل ۴-۱۷. الگوریتم پیوندی بخشی با استفاده از معیار فاصله Hamming	۱۲۵
..... شکل ۴-۱۸. الگوریتم پیوندی بخشی با استفاده از معیار فاصله Cosine	۱۲۶
..... شکل ۴-۱۹. طیفی با استفاده از ماتریس شباهت نامتراکم	۱۲۶
..... شکل ۴-۲۰. طیفی با استفاده از روش نیستروم با متعادل ساز	۱۲۷
..... شکل ۴-۲۱. طیفی با استفاده از روش نیستروم بدون متعادل ساز	۱۲۷
..... شکل ۴-۲۲. نرم افزار تحلیل گر کد استقلال الگوریتم	۱۲۸
..... شکل ۴-۲۳. ماتریس AIDM	۱۲۹
..... شکل ۴-۲۴. میانگین دقت الگوریتم های خوشبندی	۱۳۱
..... شکل ۴-۲۵. رابطه میان آستانه استقلال و زمان اجرای الگوریتم در روش پیشنهادی اول	۱۳۳
..... شکل ۴-۲۶. رابطه میان آستانه پراکندگی و زمان اجرای الگوریتم در روش پیشنهادی اول	۱۳۳

شکل ۴-۲۷. رابطه میان آستانه استقلال و دقت نتیجه نهایی در روش پیشنهادی اول ۱۳۴
شکل ۴-۲۸. رابطه میان آستانه پراکندگی و دقت نتیجه نهایی در روش پیشنهادی اول ۱۳۴
شکل ۴-۲۹. رابطه میان آستانه عدم تمرکز و دقت نتیجه نهایی در روش پیشنهادی اول ۱۳۵
شکل ۴-۳۰. رابطه میان آستانه پراکندگی و زمان اجرای الگوریتم در روش پیشنهادی دوم ۱۳۵
شکل ۴-۳۱. رابطه میان آستانه پراکندگی و دقت نتایج نهایی در روش پیشنهادی دوم ۱۳۶
شکل ۴-۳۲. رابطه میان آستانه عدم تمرکز و دقت نتایج نهایی در روش پیشنهادی دوم ۱۳۷
شکل ۴-۳۳. مقایسه زمان اجرای الگوریتم ۱۳۸

فصل اول

مقدمه

۱-۱. خوشبندی

به عنوان یکی از شاخه‌های وسیع و پرکاربرد هوش مصنوعی^۱، یادگیری ماشین^۲ به تنظیم و اکتشاف شیوه‌ها و الگوریتم‌هایی می‌پردازد که بر اساس آن‌ها رایانه‌ها و سامانه‌های اطلاعاتی توانایی تعلم و یادگیری پیدا می‌کنند. طیف پژوهش‌هایی که در مورد یادگیری ماشینی صورت می‌گیرد گسترده است. در سوی نظری آن پژوهش‌گران بر آن‌اند که روش‌های یادگیری تازه‌ای به وجود بیاورند و امکان‌پذیری و کیفیت یادگیری را برای روش‌هایی روش‌هایی را بر مسائل تازه‌ای اعمال کنند. البته این طیف گستته نیست و پژوهش‌هایی انجام‌شده دارای مؤلفه‌هایی از هر دو رویکرد هستند. امروزه، داده‌کاوی^۳ به عنوان یک ابزار قوی برای تولید اطلاعات و دانش از داده‌های خام، در یادگیری ماشین شناخته‌شده و همچنان با سرعت در حال رشد و تکامل است. به طور کلی می‌توان تکنیک‌های داده‌کاوی را به دو دسته بانظارت^۴ و بدون ناظارت^۵ تقسیم کرد [29, 46].

در روش بانظارت ما ورودی (داده یادگیری^۶) و خروجی (کلاس^۷ داده) یک مجموعه داده را به الگوریتم هوشمند می‌دهیم تا آن الگوی^۸ بین ورودی و خروجی را تشخیص دهد در این روش خروجی کار ما مدلی^۹ است که می‌تواند برای ورودی‌های جدید خروجی درست را پیش‌بینی^{۱۰} کند. روش‌های طبقه‌بندی^{۱۱} و قوانین انجمنی^{۱۲} از این جمله تکنیک‌ها می‌باشد. روش‌های با ناظارت کاربرد فراوانی دارند اما مشکل عمدۀ این روش‌ها این است که همواره باید داده‌ای برای یادگیری وجود داشته باشد که در آن به ازای ورودی مشخص خروجی درست آن مشخص شده باشد. حال آنکه اگر

¹ Artificial Intelligent (AI)

² Machine Learning

³ Data Mining

⁴ Supervised

⁵ Unsupervised

⁶ Train Set

⁷ Class

⁸ Pattern

⁹ Learning Model

¹⁰ Predictive

¹¹ Classification

¹² Association rule mining

در زمینه‌ای خاص داده‌ای با این فرمت وجود نداشته باشد این روش‌ها قادر به حل این‌گونه مسائل نخواهند بود [29, 68]. در روش بدون نظارت برخلاف یادگیری با نظارت هدف ارتباط ورودی و خروجی نیست، بلکه تنها دسته‌بندی ورودی‌ها است. این نوع یادگیری بسیار مهم است چون خیلی از مسائل (همانند دنیای ربات‌ها) پر از ورودی‌هایی است که هیچ برچسبی^{۱۳} (کلاس) به آن‌ها اختصاص داده نشده است اما به وضوح جزئی از یک دسته هستند [46, 68]. خوش‌بندی^{۱۴} شاخص‌ترین روش در داده‌کاوی جهت حل مسائل به صورت بدون ناظر است. ایده اصلی خوش‌بندی اطلاعات، جدا کردن نمونه‌ها از یکدیگر و قرار دادن آن‌ها در گروه‌های شبیه به هم می‌باشد. به این معنی که نمونه‌های شبیه به هم باید در یک گروه قرار بگیرند و با نمونه‌های گروه‌های دیگر حداقل متفاوت را دارا باشند [20, 26]. دلایل اصلی برای اهمیت خوش‌بندی عبارت‌اند از:

اول، جمع‌آوری و برچسب‌گذاری یک مجموعه بزرگ از الگوهای نمونه می‌تواند بسیار پرکاربرد و بالارزش باشد.

دوم، می‌توانیم از روش‌های خوش‌بندی برای پیدا کردن و استخراج ویژگی‌ها^{۱۵} و الگوهای جدید استفاده کنیم. این کار می‌تواند کمک به سزاپی در کشف دانش ضمنی^{۱۶} داده‌ها انجام دهد.

سوم، با خوش‌بندی می‌توانیم یک دید و بینشی از طبیعت و ساختار داده به دست آوریم که این می‌تواند برای ما بالارزش باشد.

چهارم، خوش‌بندی می‌تواند منجر به کشف زیر رده‌های^{۱۷} مجزا یا شباهت‌های بین الگوها ممکن شود که به طور چشمگیری در روش طراحی طبقه‌بندی قابل استفاده باشد.

¹³ Label

¹⁴ Clustering

¹⁵ Features

¹⁶ Tacit knowledge

¹⁷ Sub-Class

۱-۲. خوشبندی ترکیبی

هر یک از الگوریتم‌های خوشبندی، با توجه به اینکه بر روی جنبه‌های متفاوتی از داده‌ها تاکید می‌کند، داده‌ها را به صورت‌های متفاوتی خوشبندی می‌نماید. به همین دلیل، نیازمند روش‌هایی هستیم که بتواند با استفاده از ترکیب این الگوریتم‌ها و گرفتن نقاط قوت هر یک، نتایج بهینه‌تری را تولید کند. در واقع هدف اصلی خوشبندی ترکیبی^{۱۸} جستجوی بهترین خوشبدها با استفاده از ترکیب نتایج الگوریتم‌های دیگر است [1, 8, 9, 54, 56]. به روشنی از خوشبندی ترکیبی که زیرمجموعه‌ی منتخب از نتایج اولیه برای ترکیب و ساخت نتایج نهایی استفاده می‌شود خوشبندی ترکیبی مبتنی بر انتخاب^{۱۹} زیرمجموعه نتایج اولیه می‌گویند. در این روش‌ها بر اساس معیاری توافقی مجموعه‌ای از مطلوب‌ترین نتایج اولیه را انتخاب کرده و فقط توسط آن‌ها نتیجه نهایی را ایجاد می‌کنیم [21]. معیارهای مختلفی جهت انتخاب مطلوب‌ترین روش پیشنهاد شده است که معیار اطلاعات متقابل نرمال شده^{۲۰}، روش ماکریموم^{۲۱} و APMM^{۲۲} برخی از آن‌ها می‌باشند [8, 9, 21, 67]. دو مرحله مهم در خوشبندی ترکیبی عبارت‌اند از:

اول، الگوریتم‌های ابتدایی خوشبندی که خوشبندی اولیه را انجام می‌دهد.

دوم، جمع‌بندی نتایج این الگوریتم‌های اولیه (پایه) برای به دست آوردن نتیجه نهایی.

۱-۳. خرد جمعی

نظریه خرد جمعی^{۲۳} که اولین بار توسط سورویکی^{۲۴} در سال ۲۰۰۴ در کتابی با همان عنوان منتشر شد، استنباطی از مسائل مطرح شده توسط گالتون^{۲۵} و کندورست^{۲۶} می‌باشد، و نشان می‌دهد که قضاوت‌های جمعی و دموکراتیک از اعتبار بیشتری نسبت به آنچه که ما انتظار داشتیم برخوردار

¹⁸ Cluster Ensemble

¹⁹ Cluster Ensemble Selection

²⁰ Normalized Mutual Information

²¹ Maximum

²² Alizadeh-Parvin-Moshki-Minaei

²³ The wisdom of crowds

²⁴ Surowiecki

²⁵ Francis Galton (1822-1911)

²⁶ Condorcet

است، ما تأثیرات این ایده را در حل مسائل سیاسی، اجتماعی در طی سال‌های اخیر شاهد هستیم. در ادبیات خرد جمعی هر جامعه‌ای را خردمند نمی‌گویند. از دیدگاه سورویکی خردمند بودن جامعه در شرایط چهارگانه پراکندگی^{۲۷}، استقلال^{۲۸}، عدم تمرکز^{۲۹} و روش ترکیب مناسب^{۳۰} است [55].

۱-۴. خوشبندی مبتنی بر اساس نظریه خرد جمعی

هدف از این تحقیق استفاده از نظریه خرد جمعی برای انتخاب زیرمجموعه‌ی مناسب در خوشبندی ترکیبی می‌باشد. تعاریف سورویکی از خرد جمعی مطابق با مسائل اجتماعی است و در تعاریف آن عناصر سازنده تصمیمات رأی افراد می‌باشد. در این تحقیق ابتدا مبتنی بر تعاریف پایه سورویکی از خرد جمعی و ادبیات مطرح در خوشبندی ترکیبی، تعریف پایه‌ای از ادبیات خرد جمعی در خوشبندی ترکیبی ارائه می‌دهیم و بر اساس آن الگوریتم پیشنهادی خود را در جهت پیاده‌سازی خوشبندی ترکیبی ارائه می‌دهیم [55]. شرایط چهارگانه خوشبندی خردمند که مناسب با تعاریف سورویکی باز تعریف شده است به شرح زیر می‌باشد:

پراکندگی نتایج اولیه، هر الگوریتم خوشبندی پایه باید به طور جداگانه و بدون واسطه به داده‌های مسئله دسترسی داشته و آن را تحلیل و خوشبندی کند حتی اگر نتایج آن غلط باشد.

استقلال الگوریتم، روش تحلیل هر یک از خوشبندی‌های پایه نباید تحت تأثیر روش‌های سایر خوشبندی‌های پایه تعیین شود، این تأثیر می‌تواند در سطح نوع الگوریتم (گروه) یا پارامترهای اساسی یک الگوریتم خاص (افراد) باشد.

عدم تمرکز، ارتباط بین بخش‌های مختلف خوشبندی خرد جمعی باید به گونه‌ای باشد تا بر روی عملکرد خوشبندی پایه تأثیری ایجاد نکند تا از این طریق هر خوشبندی پایه شанс این را داشته باشد تا با شخصی سازی و بر اساس دانش محلی خود بهترین نتیجه ممکن را آشکار سازد.

²⁷ Diversity

²⁸ Independence

²⁹ Decentralization

³⁰ Aggregation Mechanism

مکانیزم ترکیب مناسب، باید مکانیزمی وجود داشته باشد که بتوان توسط آن نتایج اولیه الگوریتم‌های پایه را با یکدیگر ترکیب کرده و به یک نتیجه نهایی (نظر جمعی) رسید.

در این تحقیق دو روش برای ترکیب خوشبندی ترکیبی و خرد جمعی پیشنهاد شده است. با استفاده از تعاریف بالا الگوریتم روش اول مطرح خواهد شد که در آن، جهت رسیدن به نتیجه نهایی از آستانه‌گیری استفاده می‌شود. در این روش الگوریتم‌های خوشبندی اولیه غیر هم نام کاملاً مستقل فرض خواهند شد و برای ارزیابی استقلال الگوریتم‌های هم نام نیاز به آستانه‌گیری می‌باشد. در روش دوم، سعی شده است تا دو بخش از روش اول بهبود یابد. از این روی جهت مدل‌سازی الگوریتم‌ها و ارزیابی استقلال آنها نسبت به هم یک روش مبتنی بر گراف شبکه که ارائه می‌شود و میزان استقلال به دست آمده در این روش به عنوان وزنی برای ارزیابی پراکندگی در تشکیل جواب نهایی مورد استفاده قرار می‌گیرد. جهت ارزیابی، روش‌های پیشنهادی با روش‌های پایه، روش ترکیب کامل و چند روش معروف ترکیب مبتنی بر انتخاب مقایسه خواهد شد. از این روی از چهارده داده استاندارد و یا مصنوعی که عموماً از سایت UCI [76] جمع‌آوری شده‌اند استفاده شده است. در انتخاب این داده‌ها سعی شده، داده‌هایی با مقیاس کوچک، متوسط و بزرگ انتخاب شوند تا کارایی روش بدون در نظر گرفتن مقیاس داده ارزیابی شود. همچنین جهت اطمینان از صحت نتایج تمامی آزمایش‌های تجربی گزارش شده حداقل ده بار تکرار شده است.

۱-۴-۱- فرضیات تحقیق

این تحقیق بر اساس فرضیات زیر اقدام به ارائه روشی جدید در خوشبندی ترکیبی مبتنی بر انتخاب بر اساس نظریه خرد جمعی می‌کند.

۱) در این تحقیق تمامی آستانه‌گیری‌ها بر اساس میزان صحت نتایج نهایی و مدت زمان اجرای الگوریتم به صورت تجربی انتخاب می‌شوند.

۲) در این تحقیق جهت ارزیابی عملکرد یک الگوریتم، نتایج اجرای آن را بر روی داده‌های استاندارد UCI در محیطی با شرایط و پارامترهای مشابه نسبت به سایر الگوریتم‌ها ارزیابی می‌کنیم که این داده‌ها الزاماً حجمی یا خیلی کوچک نیستند.

۳) جهت اطمینان از صحت نتایج آزمایش‌ها ارائه شده در این تحقیق، حداقل اجرای هر الگوریتم بر روی هر داده ده بار تکرار شده و نتیجه‌ی نهایی میانگین نتایج به دست آمده می‌باشد.

۴) از آنجایی که روش مطرح شده در این تحقیق یک روش مکاشفه‌ای است سعی خواهد شد بیشتر با روش‌های مکاشفه‌ای مطرح در خوشه‌بندی ترکیبی مقایسه و نتایج آن مورد بررسی قرار گیرد.

در این فصل اهداف، مفاهیم و چالش‌های این تحقیق به صورت خلاصه ارائه شد. در ادامه این تحقیق، در فصل دوم، الگوریتم‌های خوشه‌بندی پایه و روش‌های خوشه‌بندی ترکیبی مورد بررسی قرار می‌گیرد. همچنین به مرور روش‌های انتخاب خوشه^{۳۱} و یا افزار^{۳۲} در خوشه‌بندی ترکیبی مبتنی بر انتخاب خواهیم پرداخت. در فصل سوم، نظریه خرد جمعی و دو روش پیشنهادی خوشه‌بندی خردمند ارائه می‌شود. در فصل چهارم، به ارائه نتایج آزمایش‌های تجربی این تحقیق و ارزیابی آن‌ها می‌پردازیم و در فصل پنجم، به ارائه نتایج و کارهای آتی خواهیم پرداخت.

³¹ Cluster

³² Partition

فصل دوم

مروری بر ادبیات

تحقیق

۲. مروری بر ادبیات تحقیق

۱-۱. مقدمه

در این بخش، کارهای انجام شده در خوشبندی و خوشبندی ترکیبی را مورد مطالعه قرار می‌دهیم. ابتدا چند الگوریتم پایه خوشبندی معروف را معرفی خواهیم کرد. سپس چند روش کاربردی جهت ارزیابی خوش، خوشبندی و افزایشندی را مورد مطالعه قرار می‌دهیم. در ادامه به بررسی ادبیات خوشبندی ترکیبی خواهیم پرداخت و روش‌های ترکیب متداول را بررسی خواهیم کرد. از روش‌های خوشبندی ترکیبی، روش ترکیب کامل و چند روش معروف مبتنی بر انتخاب را به صورت مفصل شرح خواهیم داد.

۱-۲. خوشبندی

در این بخش ابتدا انواع الگوریتم‌های خوشبندی پایه را معرفی می‌کنیم و سپس برخی از آن‌ها را مورد مطالعه قرار می‌دهیم سپس برای ارزیابی نتایج به دست آمده چند متریک معرفی خواهیم کرد.

۱-۲-۱. الگوریتم‌های خوشبندی پایه

به طور کلی، الگوریتم‌های خوشبندی را می‌توان به دو دسته کلی تقسیم کرد:

۱- الگوریتم‌های سلسله مراتبی^{۳۳}

۲- الگوریتم‌های افزایشندی^{۳۴}

الگوریتم‌های سلسله مراتبی، یک روال برای تبدیل یک ماتریس مجاورت به یک دنباله از افزاهای تو در تو، به صورت یک درخت است. در این روش‌ها، مستقیماً با داده‌ها سروکار داریم و از روابط بین آن‌ها برای به دست آوردن خوش‌ها استفاده می‌کنیم. یکی از ویژگی‌های این روش قابلیت تعیین تعداد خوش‌ها به صورت بهینه می‌باشد. در نقطه مقابل الگوریتم‌های سلسله مراتبی، الگوریتم‌های

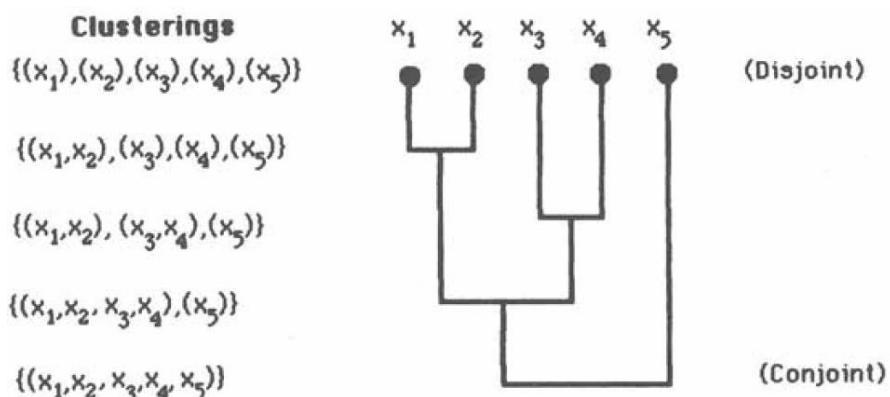
³³ Hierarchical

³⁴ Partitioning

افرازبندی قرار دارند. هدف این الگوریتم‌ها، تقسیم داده‌ها در خوش‌های، به گونه‌ای است که داده‌های درون یک خوش‌ه بیشترین شباهت را به همدیگر داشته باشند؛ و در عین حال، بیشترین فاصله و اختلاف را با داده‌های خوش‌های دیگر داشته باشند. در این فصل تعدادی از متدالول ترین الگوریتم‌های خوش‌بندی، در دو دسته سلسله مرتبی و افرازبندی، مورد بررسی قرار می‌گیرند. از روش سلسله مرتبی چهار الگوریتم از سری الگوریتم‌های پیوندی^{۳۵} را مورد بررسی قرار می‌دهیم. و از الگوریتم‌های افرازبندی FCM و الگوریتم طیفی را مورد بررسی خواهیم داد.

۱-۲-۲. الگوریتم‌های سلسله مرتبی

همان‌گونه که در شکل ۱-۲ مشاهده می‌شود، روال الگوریتم‌های خوش‌بندی سلسله مرتبی را می‌تواند به صورت یک دندوگرام^{۳۶} نمایش داد. این نوع نمایش تصویری از خوش‌بندی سلسله مرتبی، برای انسان، بیشتر از یک لیست از نمادها قابل درک است. در واقع دندوگرام، یک نوع خاص از ساختار درخت است که یک تصویر قابل فهم از خوش‌بندی سلسله مرتبی را ارائه می‌کند. هر دندوگرام شامل چند لایه از گره‌های است، به طوری که هر لایه یک خوش‌ه را نمایش می‌دهد. خطوط متصل‌کننده گره‌ها، بیانگر خوش‌هایی هستند که به صورت آشیانه‌ای^{۳۷} داخل یکدیگر قرار دارند. برش افقی یک دندوگرام، یک خوش‌بندی را تولید می‌کند [33]. شکل ۱-۲ یک مثال ساده از خوش‌بندی و دندوگرام مربوطه را نشان می‌دهد.



شکل ۱-۲. یک خوش‌بندی سلسله مرتبی و درخت متناظر

³⁵ Linkage

³⁶ Dendogram

³⁷ Nested

اگر الگوریتم‌های خوشبندی سلسله مراتبی، دندوگرام را به صورت پایین به بالا بسازند، الگوریتم‌های خوشبندی سلسله مراتبی تراکمی^{۳۸} نامیده می‌شوند. همچنین، اگر آن‌ها دندوگرام را به صورت بالا به پایین بسازند، الگوریتم‌های خوشبندی سلسله مراتبی تقسیم‌کننده^{۳۹} نامیده می‌شوند [26]. مهم‌ترین روش‌های خوشبندی سلسله مراتبی الگوریتم‌های سری پیوندی می‌باشد که در این بخش تعدادی از کاراترین آن‌ها مورد بررسی قرار خواهند گرفت که عبارت‌اند از:

۱- الگوریتم پیوندی منفرد^{۴۰}

۲- الگوریتم پیوندی کامل^{۴۱}

۳- الگوریتم پیوندی میانگین^{۴۲}

۴- الگوریتم پیوندی بخشی^{۴۳}

۱-۱-۱-۲-۲. تعاریف و نمادها

$$\mathcal{D}_1 = \begin{matrix} & x_1 & x_2 & x_3 & x_4 & x_5 \\ x_1 & \left[\begin{matrix} 0 & 6 & 8 & 2 & 7 \\ 6 & 0 & 1 & 5 & 3 \\ 8 & 1 & 0 & 10 & 9 \\ 2 & 5 & 10 & 0 & 4 \\ 7 & 3 & 9 & 4 & 0 \end{matrix} \right] \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{matrix}$$

شکل ۲-۲. ماتریس مجاورت

قبل از معرفی این الگوریتم‌ها، در ابتدا نمادها و نحوه نمایش مسئله نمایش داده خواهد شد. فرض کنید که یک ماتریس مجاورت $n \times n$ متقارن $D = [d(i, j)]_{n(n-1)/2}$ داریم. n واردہ در هر سمت قطر اصلی قرار دارد که شامل یک جای گشت اعداد صحیح بین ۱ تا $n(n-1)/2$ است. ما مجاورت‌ها را عدم شباهت در نظر می‌گیریم. $d(1,2) > d(1,3)$ به این معنی است که اشیاء ۱ و ۳

³⁸ Agglomerative

³⁹ Divisive

⁴⁰ Single Linkage

⁴¹ Complete Linkage

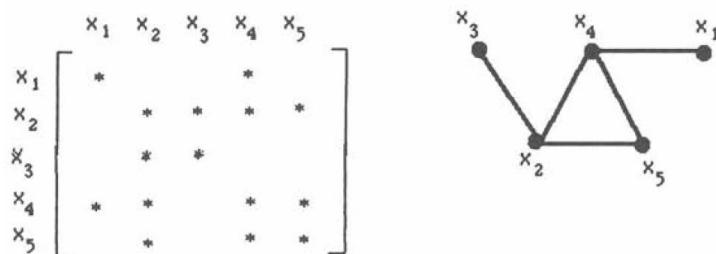
⁴² Average Linkage

⁴³ Ward Linkage

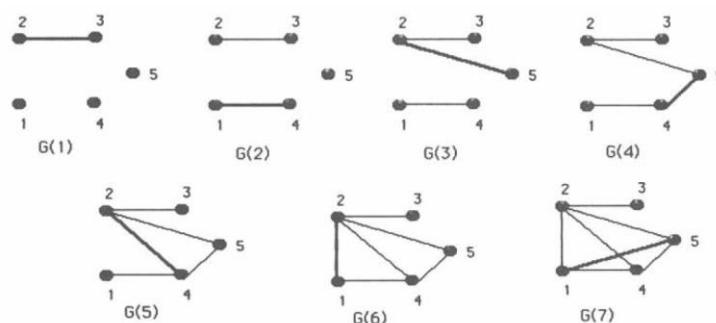
بیشتر از اشیاء ۱ و ۲ به هم شبیه‌اند. D_1 یک مثال از ماتریس مجاورت معمول برای $n=5$ است که در شکل ۲-۲ نشان داده شده است. یک گراف آستانه^{۴۴}، یک گراف غیر جهت‌دار و غیر وزن‌دار، روی N گره، بدون حلقه بازگشت به خود^{۴۵} یا چند لبه است. هر نود یک شیء را نمایش می‌دهد. یک گراف آستانه $G(v)$ برای هر سطح عدم شباهت v به این صورت تعریف می‌شود: اگر عدم شباهت اشیاء i و j از حد آستانه v کوچک‌تر باشد، با وارد کردن یک لبه (j, i) بین نودهای i و j یک گراف آستانه تعریف می‌کنیم.

$$(i, j) \in G(v) \text{ if and only if } d(i, j) \leq v \quad (1-2)$$

شکل ۲-۳ یک رابطه دودویی به دست آمده از ماتریس D_1 مربوط به شکل ۲-۲ را برای مقدار آستانه ۵ نشان می‌دهد. نماد "*" در موقعیت (i, j) ماتریس، نشان می‌دهد که جفت (x_i, x_j) متعلق به رابطه دودویی می‌باشد. شکل ۴-۲، گراف‌های آستانه برای ماتریس D_1 را نمایش می‌دهد.



شکل ۲-۳. رابطه دودویی و گراف آستانه برای مقدار آستانه ۵.



شکل ۴-۲. گراف‌های آستانه برای ماتریس D_1

⁴⁴ Threshold graph

⁴⁵ Self-loop

۲-۱-۱-۲-۲. الگوریتم پیوندی منفرد

این الگوریتم روش کمینه و روش نزدیکترین همسایه نیز نامیده می‌شود [26]. اگر C_i و C_j خوشه‌ها باشند، در روش پیوندی منفرد، فاصله آن‌ها برابر خواهد بود با:

$$D_{SL}(C_i, C_j) = \min_{a \in c_i, b \in c_j} d(a, b) \quad (2-2)$$

که $d(a, b)$ نشان‌دهنده فاصله (عدم شباهت) بین نقاط a و b در ماتریس مجاورت است. شکل ۲-۲ این الگوریتم را نمایش می‌دهد. شکل ۲-۶ دندوگرام حاصل از روش پیوندی منفرد را برای ماتریس D_1 نشان می‌دهد.

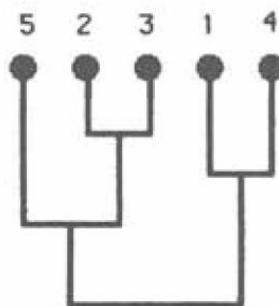
Step 1. Begin with the disjoint clustering implied by threshold graph $G(0)$, which contains no edges and which places every object in a unique cluster, as the current clustering. Set $k \leftarrow 1$.

Step 2. From threshold graph $G(k)$.

If the number of components (maximally connected subgraphs) in $G(k)$, is less than the number of clusters in the current clustering, redefine the current clustering by naming each component of $G(k)$ as a cluster.

Step 3. If $G(k)$ consists of a single connected graph, stop. Else, set $k \leftarrow k + 1$ and go to step 2.

شکل ۲-۵. الگوریتم خوشه‌بندی سلسله مراتبی تراکمی پیوندی منفرد



Single Link

شکل ۲-۶. دندوگرام پیوندی منفرد برای ماتریس D_1

۲-۱-۱-۲-۳. الگوریتم پیوندی کامل

این الگوریتم روش بیشینه یا روش دورترین همسایه نیز نامیده می‌شود. الگوریتم پیوندی کامل می‌گوید که وقتی دو خوشه i و j شبیه به هم هستند که بیشینه $D_{cl}(C_i, C_j)$ روی تمام a ها

در C_i و C_j کوچک باشد. به عبارت دیگر، در این الگوریتم، برای یکی کردن دو خوش، همه جفت‌ها در دو خوش باید شیبیه به هم باشند [26]. اگر C_i و C_j خوش‌ها باشند، در روش پیوندی کامل، فاصله آن‌ها برابر خواهد بود با:

$$D_{cl}(C_i, C_j) = \max_{a \in C_i, b \in C_j} d(i, j) \quad (3-2)$$

که $d(a, b)$ نشان‌دهنده فاصله (عدم شباهت) بین نقاط a و b در ماتریس مجاور است. شکل ۷-۲ این الگوریتم و شکل ۸-۲ دندوگرام حاصل از این روش را برای ماتریس D_1 ، را نشان می‌دهد.

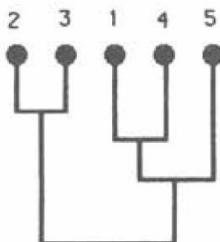
Step 1. Begin with the disjoint clustering implied by threshold graph $G(0)$, which contains no edges and which places every object in a unique cluster, as the current clustering. Set $k \leftarrow 1$.

Step 2. From threshold graph $G(k)$.

If two of the current clusters from a clique (maximally complete sub graph) in $G(k)$, redefine the current clustering by merging these two clusters into a single cluster.

Step 3. If $k = n(n - 1)/2$, so that $G(k)$ is the complete graph on the n nodes, stop. Else, set $k \leftarrow k + 1$ and go to step 2.

شکل ۷-۲. الگوریتم خوش‌بندی سلسله مراتبی تراکمی پیوندی کامل



Complete Link

شکل ۸-۲ دندوگرام پیوندی کامل برای ماتریس D_1

۴-۱-۲-۲-۴. الگوریتم پیوندی میانگین

الگوریتم پیوندی منفرد اجازه می‌دهد تا خوش‌ها به صورت دراز و نازک رشد کنند. این در شرایطی است که الگوریتم پیوندی کامل خوش‌های فشرده‌تری تولید می‌کند. هر دو الگوریتم مستعد خطای داده‌های خارج از محدوده^{۴۶} هستند. الگوریتم خوش‌بندی پیوندی میانگین، یک تعادلی بین

⁴⁶ Outliers

مقادیر حدی الگوریتم‌های پیوندی منفرد و کامل است. الگوریتم پیوندی میانگین همچنین، روش جفت-گروه بدون وزن با استفاده از میانگین حسابی^{۴۷} نامیده می‌شود. این الگوریتم، یکی از پرکاربردترین الگوریتم‌های خوشبندی سلسله مراتبی می‌باشد [26]. اگر C_i یک خوشه با تعداد n_i تا عضو، و C_j یک خوشه دیگر با تعداد n_j تا عضو باشند، در روش پیوندی میانگین، فاصله آن‌ها برابر خواهد بود با:

$$D_{AL}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{a \in n_i, b \in n_j} d(a, b) \quad (4-2)$$

که $d(a, b)$ نشان‌دهنده فاصله (عدم شباهت) بین نقاط a و b در ماتریس مجاورت است.

۱-۱-۲-۲. الگوریتم پیوندی بخشی

روش پیوندی بخشی که از مربع مجموع خطاهای (SSE) خوشبها ای یک افزای برای ارزیابی استفاده می‌کند، یکی دیگر از روش‌های سلسله مراتبی می‌باشد [60]. اگر C_i یک خوشه با تعداد n_i تا عضو، و C_j یک خوشه دیگر با تعداد n_j تا عضو باشند و نماد $\|\cdot\|_2$ به معنای فاصله اقلیدسی و \bar{a} و \bar{b} مرکز خوشبها ای C_i و C_j باشد آنگاه در روش پیوندی بخشی، فاصله آن‌ها برابر خواهد بود با:

$$D_{WL}(C_i, C_j) = \sqrt{\frac{2n_i n_j}{(n_i, n_j)} \|\bar{a} - \bar{b}\|_2} \quad (5-2)$$

۲-۱-۲-۲. الگوریتم‌های افزایبندی

یک خاصیت مهم روش‌های خوشبندی سلسله مراتبی، قابلیت نمایش دندوگرام است که تحلیل‌گر را قادر می‌سازد تا بیند که چگونه اشیاء در سطوح متوالی مجاورت، در خوشبها به هم پیوند می‌خورند یا تفکیک می‌شوند. همان طور که اشاره شد، هدف الگوریتم‌های افزایبندی، تقسیم داده‌ها در خوشبها، به گونه‌ای است که داده‌های درون یک خوشه بیشترین شباهت را به هم‌دیگر داشته باشند؛ و در عین حال، بیشترین فاصله و اختلاف را با داده‌های خوشبها دیگر داشته باشند.

⁴⁷ Un-weighted Pair-Group Method using Arithmetic Averages (UPGMA)

آن‌ها یک افزار منفرد از داده را تولید می‌کنند و سعی می‌کنند تا گروه‌های طبیعی حاضر در داده را کشف کنند. هر دو رویکرد خوش‌بندی، دامنه‌های مناسب کاربرد خودشان را دارند. معمولاً روش‌های خوش‌بندی سلسله مراتبی، نیاز به ماتریس مجاورت بین اشیاء دارند؛ در حالی که روش‌های افزاربندی، به داده‌ها در قالب ماتریس الگو نیاز دارند. نمایش رسمی مسئله خوش‌بندی افزاربندی می‌تواند به صورت زیر باشد:

تعیین یک افزار از الگوهای در K گروه، یا خوش، با داشتن n الگو در یک فضای d -بعدی؛ به طوری که الگوهای در یک خوش بیشترین شباهت را به هم داشته و با الگوهای خوش‌های دیگر بیشترین، تفاوت را داشته باشند. تعداد خوش‌های K ، ممکن است که از قبل مشخص شده نباشد، اما در بسیاری از الگوریتم‌های خوش‌بندی افزاربندی، تعداد خوش‌های باید از قبل معلوم باشند. در ادامه برخی از معروف‌ترین و پرکاربردترین الگوریتم‌های افزاربندی مورد بررسی قرار خواهند گرفت.

۱-۲-۱-۲-۲. الگوریتم K-means

در الگوریتم K -means مراکز خوش‌های بلا فاصله بعد از اینکه یک نمونه به یک خوش می‌پیوندد محاسبه می‌شوند. به طور معمول بیشتر روش‌های خوش‌بندی ترکیبی از الگوریتم K -means جهت خوش‌بندی اولیه خود استفاده می‌کنند [37, 47, 57]. اما مطالعات اخیر نشان داده‌اند که با توجه به رفتار هر مجموعه داده، گاهی اوقات یک روش خوش‌بندی خاص پیدا می‌شود که دقیق‌تر از K -means برای بعضی از مجموعه داده‌ها می‌دهد [1, 54]. اما الگوریتم K -means به دلیل سادگی و توانایی مناسب در خوش‌بندی همواره به عنوان انتخاب اول مطالعات خوش‌بندی ترکیبی مورد مطالعه قرار گرفته است. در شکل ۱۰-۲ شبه کد الگوریتم K -means را مشاهده می‌کنید:

-
1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
 2. Assign each object to the group that has the closest centroid.
 3. When all objects have been assigned, recalculate the positions of the K centroids.
 4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated

شکل ۹-۲. الگوریتم خوش‌بندی افزاربندی K -means

مقادیر مراکز اولیه‌ی متفاوت برای الگوریتم K -means می‌تواند منجر به خوش‌بندی‌های مختلفی شود. به خاطر اینکه این الگوریتم مبتنی بر مربع خطأ است، می‌تواند به کمینه محلی همگرا شود،

مخصوصاً برای خوش‌هایی که به طور خیلی خوبی از هم تفکیک نمی‌شوند، این امر صادق است. نشان داده شده است که هیچ تضمینی برای همگرایی یک الگوریتم تکراری به یک بهینه سراسری نیست [33]. به طور خلاصه می‌توان ویژگی‌های الگوریتم K -means را به صورت زیر برشمرد:

- ۱- بر اساس فاصله اقلیدسی تمامی ویژگی‌ها می‌باشد.
- ۲- منجر به تولید خوش‌هایی به صورت دایره، کره و یا ابر کره می‌شود.
- ۳- نسبت به روش‌های دیگر خوش‌بندی، ساده و سریع است.
- ۴- همگرایی آن به یک بهینه محلی اثبات شده است، اما تضمینی برای همگرایی به بهینه سراسری وجود ندارد.
- ۵- نسبت به مقداردهی اولیه مراکز خوش‌ها خیلی حساس است.

۲-۲-۱-۲-۲. الگوریتم FCM

الگوریتم FCM^{48} اولین بار توسط دون [13] ارائه شد. سپس توسط بزدک [66] بهبود یافت. این متند دیدگاه جدیدی را در خوش‌بندی بر اساس منطق فازی [62] ارائه می‌دهد. در این دیدگاه جدید، به جای اینکه داده‌ها در یک خوش‌ه عضو باشند، در تمامی خوش‌ها با یک ضریب عضویت که بین صفر و یک است، عضو هستند و ما در این نوع خوش‌بندی، دنبال این ضرایب هستیم. در روش‌های معمول در جایی که ما N داده داشته باشیم، جواب نهایی ماتریس $N \times 1$ خواهد بود که هر خانه شامل برچسب خوش‌ی داده‌ی نظیر آن می‌باشد. ولی در این روش در صورت داشتن k خوش‌ه، جواب نهایی یک ماتریس $N \times k$ خواهد بود که در آن هر ردیف شامل ضرایب عضویت داده‌ی نظیر به آن خوش‌ه است. بدیهی است که جمع افقی هر ردیف (ضرایب عضویت یک داده خاص) برابر با یک خواهد بود. یک روش معمول جهت رسیدن به جواب‌هایی غیر فازی بر اساس نتایج نهایی الگوریتم فازی، برچسب‌زنی داده بر اساس آن ضریبی که مقدار حداقل‌تر را در این داده دارد، می‌باشد. رابطه ۲-۶ معادله پایه در روش فازی است: [66]

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2, \quad 1 \leq m < \infty \quad (6-2)$$

⁴⁸ Fuzzy c-means

در رابطه ۶-۲ متغیر m یک عدد حقیقی بزرگ‌تر از یک و u_{ij} درجه عضویت داده x_i در خوشة زمان می‌باشد، که خود x_i ، i -امین داده d -بعدی از داده‌ی مورد مطالعه می‌باشد و c_j مرکز d -بعدی خوشه زمان است و $\|*$ هر روش معمول جهت اندازه‌گیری شباهت میان داده x_i و مرکز خوشه c_j می‌باشد. در روش خوشه‌بندی فازی مراکز خوشه (c_j) و درجه عضویت (u_{ij}) با تکرار مکرر به ترتیب بر اساس رابطه‌های ۷-۲ و ۸-۲ به روزرسانی می‌شوند، تا زمانی که شرط توقف درست در آید. در این شرط مقدار ϵ یک مقدار توافقی بسیار کوچک‌تر از یک می‌باشد که مطابق با نوع داده و دقت خوشه‌بندی قابل جایگذاری خواهد بود. بدیهی است که هر چقدر این مقدار به سمت صفر میل کند درجه عضویت دقیق‌تر و مقدار زمان اجرا بیشتر خواهد بود [66].

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m \times x_i}{\sum_{i=1}^N u_{ij}^m} \quad (7-2)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad (8-2)$$

مراحل اجرای الگوریتم FCM در شبه کد شکل ۱۱-۲ شرح داده شده است:

1. Initialize $U = [u_{ij}]$ matrix, $U^{(0)}$

2. At k-step: calculate the centers vectors $C^{(k)} = [c_j]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m \times x_i}{\sum_{i=1}^N u_{ij}^m}$$

3. Update $U^{(k)}, U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

4. If $\|U^{(k+1)} - U^{(k)}\| < \epsilon$ then STOP; otherwise return to step 2.

شکل ۱۰-۲. الگوریتم فازی خوشه‌بندی FCM

۳-۲-۱-۲. الگوریتم طیفی

روش خوشبندی طیفی^{۴۹} که بر اساس مفهوم گراف طیفی [11] مطرح شده است، از ماتریس شباهت برای کاهش بعد داده‌ها در خوشبندی استفاده می‌کند. در این روش یک گراف وزن‌دار بدون جهت به نحوی تولید می‌شود که رئوس گراف نشان‌دهنده‌ی مجموعه نقاط و هر یال وزن‌دار نشان‌دهنده‌ی میزان شباهت جفت داده‌های متناظر باشد. بر خلاف روش‌های کلاسیک، این روش، روی داده‌ای پراکنده در فضایی با شکل هندسی غیر محدب، نتایج مطلوبی تولید می‌کند [63]. کاربرد این روش در محاسبات موازی^{۵۰} [69, 70]، تنظیم بار^{۵۱} [15]، طراحی VLSI^{۵۲} [28]، طبقه‌بندی تصاویر^{۵۳} [35] و بیوانفورماتیک^{۵۴} [31, 59] می‌باشد.

در خوشبندی طیفی از بردارهای ویژگی در ماتریس شباهت برای افزایش مجموعه داده استفاده می‌شود. در اغلب این روش‌ها، مقدار ویژه اولویت بردارها را تعیین می‌کند. ولی این نحوه انتخاب، انتخاب بهترین بردارها را تضمین نمی‌دهد. در اولین تحقیقی که در این زمینه توسط ژیانگ و گنگ [61] انجام شد، مسئله انتخاب بردارهای ویژگی مناسب جهت بهبود نتایج خوشبندی پیشنهاد گردید. در روش پیشنهادی آن‌ها شایستگی هر یک از بردارهای با استفاده ازتابع چگالی احتمال هر بردار تخمین زده می‌شود. وزنی به بردارهایی که امتیاز لازم را به دست آورندگ، اختصاص یافته و برای خوشبندی از آن‌ها استفاده می‌شود. در کاری دیگر که توسط ژائو [64] انجام شده است، هر یک از بردارهای ویژه به ترتیب حذف می‌شوند و مقدار آنتروپی مجموعه بردارهای باقی‌مانده محاسبه می‌شود. برداری که حذف آن منجر به افزایش آنتروپی و ایجاد بی‌نظمی بیشتر در مجموعه داده شود، اهمیت بیشتری داشته و در رتبه بالاتری قرار می‌گیرد. سپس زیرمجموعه‌های از مناسب‌ترین بردارها برای خوشبندی مورد استفاده قرار می‌گیرند. الگوریتم خوشبندی طیفی دارای متدهای متفاوتی جهت پیاده‌سازی است، که الگوریتم‌های برش نرمال، NJW^{5۵} و PF^{5۶} از آن جمله می‌باشد. در تمامی این روش‌ها، بخش اول، یعنی تولید گراف، مشترک می‌باشد. ما در ادامه ابتدا به بررسی بخش مشترک این روش‌ها می‌پردازیم. سپس به تشریح دو روش پر کاربرد برش نرمال و NJW می‌پردازیم.

⁴⁹ Spectral

⁵⁰ Parallel processing

⁵¹ Load balancing

⁵² Very Large Scale Integration

⁵³ Image segmentation

⁵⁴ Bioinformatic

در الگوریتم خوشه‌بندی طیفی، افزای داده‌ها بر اساس تجزیه‌ی ماتریس شباهت و به دست آوردن بردارها و مقادیر ویژه‌ی آن صورت می‌گیرد. مجموعه‌ی $\{x_1, x_2, \dots, x_n\}$ با N داده‌ی d بعدی را در نظر بگیرید، می‌توان برای این مجموعه گراف وزن‌دار و بدون جهت $G(V, A)$ را ساخت به صورتی که رئوس گراف $V = \{v_1, v_2, \dots, v_n\}$ نشان‌دهنده N داده و یال‌ها که ماتریس شباهت $N \times N$ را تشکیل می‌دهند بیانگر میزان شباهت بین هر جفت داده متناظر باشند. ماتریس شباهت به صورت رابطه ۹-۲ تعریف می‌شود:

$$A_{ij} = \begin{cases} h(x_i, x_j) & i \neq j \\ 0 & \text{otherwise} \end{cases} \quad (9-2)$$

تابع h میزان شباهت بین دو داده را اندازه می‌گیرد. می‌تواند یک تابع گوسی به صورت $h(x_i, x_j) = \exp(-d^2(x_i, x_j)/\sigma^2)$ باشد. که در آن d فاصله‌ی بین دو نمونه را نشان می‌دهد و پارامتر مقیاس σ سرعت کاهش تابع h با افزایش فاصله بین دو نمونه را مشخص می‌کند. در ادامه به بررسی دو الگوریتم خوشه‌بندی طیفی برش نرمال و NJW می‌پردازیم.

۱-۲-۱-۲-۲-۳-۲-۱. الگوریتم برش نرمال

الگوریتم برش نرمال^{۵۵} توسط شی و ملیک [35] برای قطعه‌بندی تصاویر ارائه شده است. در این روش، میزان تفاوت بین خوشه‌های مختلف و شباهت بین اعضا یک خوشه، بر اساس فاصله‌ی داده‌ها محاسبه می‌کند. رابطه ۱۰-۲ اشاره به مفهوم شباهت داده دارد که با استفاده از آن اقدام به ساخت گراف وزن‌دار $G(V, A)$ می‌نماییم:

$$a_{ij} = \exp\left(-\frac{\|f_i - f_j\|^2}{2\sigma_1^2}\right) \times \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_2^2}\right) & \text{if } \|x_i - x_j\| < R \\ 0 & \text{otherwise.} \end{cases} \quad (10-2)$$

x_i موقعیت i -امین داده (پیکسل در تصاویر) و f_i بردار ویژگی از صفات داده (مانند روش‌نایی در تصاویر) می‌باشد. با کمک حد آستانه R می‌توان میزان تنکی ماتریس شباهت را با توجه به تعداد اثرگذار داده‌های همسایه تعیین کرد. گام‌های این الگوریتم به صورت زیر می‌باشد:

⁵⁵ Normal Cut

۱- محاسبه ماتریس درجه D .

۲- محاسبه ماتریس لاپلاسین $.L_N = D^{-1/2}AD^{-1/2}$

۳- محاسبه دومین بردار ویژگی متناظر با دومین کوچکترین مقدار ویژه λ_2 .

۴- استفاده از $D^{-1/2}\ell_2$ برای خوشبندی (قطعه‌بندی در تصاویر) گراف G .

روش برش نرمال بیشتر در قطعه‌بندی تصاویر کاربرد دارد و عموماً در خوشبندی داده از سایر الگوریتم‌های خوشبندی طیفی استفاده می‌کنند.

۲-۱-۲-۲-۳-۲. الگوریتم NJW

ایده الگوریتم NJW استفاده از اولین k بردار ویژه متناظر با بزرگترین k مقدار ویژه ماتریس لاپلاسین است. مراحل این الگوریتم به صورت زیر می‌باشد: [51]

۱- ساخت ماتریس شباهت با استفاده از رابطه ۹-۲.

۲- محاسبه ماتریس درجه D ، $D_{ij} = \sum_{j=1}^n A_{ij}$ و ماتریس لاپلاسین L_N .

۳- به دست آوردن اولین k بردار ویژه v^1, v^2, \dots, v^k متناظر با اولین k بزرگترین مقدار ماتریس

$.V = [v^1, v^2, \dots, v^k] \in R^{n \times k}$ و تشکیل ماتریس ستونی L_N

۴- نرمال سازی مجدد V و تشکیل Y به طوری که همه سطرهای آن طول واحد داشته باشد

$$. Y_{ij} = \frac{V_{ij}}{\sqrt{\sum_j V_{ij}^2}}$$

۵- خوشبندی مجموعه داده بازنمایی شده Y با استفاده از $K-means$

۴-۲-۱-۲-۲. الگوریتم خوشبندی کاهشی

الگوریتم خوشبندی کاهشی^{۵۶} یکی از سریع‌ترین الگوریتم‌های تک گذر، برای تخمین تعداد خوشه و مراکز آنها در مجموعه داده می‌باشد. این مفهوم یعنی به جای تحت تأثیر قرار گرفتن محاسبات از ابعاد مسئله، متناسب با اندازه مسئله آن را انجام دهیم. با این وجود، مراکز واقعی خوشه‌الزاماً یکی از نقاط داده موجود در مجموعه داده نیست ولی در بیشتر موارد این انتخاب تخمین خوبی است که به صورت ویژه از این رویکرد در محاسبات کاهشی استفاده می‌شود. اگر هر نقطه از مجموعه داده به عنوان گزینه‌ای برای مرکز خوشه در نظر گرفته شود، معیار تراکم هر نقطه X_i به صورت زیر تعریف می‌شود [79].

$$D_i = \sum_{j=1}^n \exp\left(-\frac{\|X_i - X_j\|^2}{\left(r_a/2\right)^2}\right) \quad (11-2)$$

در رابطه بالا r_a یک ثابت مثبت است، که نشان‌دهنده شعاع همسایگی^{۵۷} (سایر نقاط داده که نزدیک‌ترین نقاط به این داده خاص هستند) می‌باشد، و r_a نشان‌دهنده سایر داده‌های مجموعه، و n نشان‌دهنده تعداد این داده‌ها است. از این روی، داده‌ای دارای بیش‌ترین مقدار تراکم می‌باشد که بیش‌ترین نقاط داده در همسایگی آن است. اولین مرکز خوشه X_{c_1} بر اساس بزرگ‌ترین مقدار تراکم D_{c_1} انتخاب می‌شود. بعد از این انتخاب میزان تراکم هر یک از نقاط داده X_i به صورت زیر به‌روز می‌شود [79].

$$D_i = D_i - D_{c_1} \exp\left(-\frac{\|X_i - X_{c_1}\|^2}{\left(r_b/2\right)^2}\right) \quad (12-2)$$

در رابطه بالا ثابت مثبت r_b همسایگی را تعریف می‌کند که میزان کاهش تراکم قابل اندازه‌گیری را نشان می‌دهد. از آنجایی که نقاط داده در نزدیکی مرکز خوشه اول X_{c_1} به طور قابل توجهی مقادیر چگالی را کاهش می‌دهند بعد از به‌روز کردن مقادیر تابع چگالی توسط رابطه بالا مرکز خوشه بعدی

⁵⁶ Subtractive Clustering

⁵⁷ Neighborhood radius

بر اساس داده‌ای که بزرگ‌ترین مقدار چگالی را دارد انتخاب می‌شود. این فرآیند آن قدر تکرار می‌شود تا به تعداد کافی مرکز خوش‌بندی شود. پس از اتمام این فرآیند می‌توان توسط الگوریتم K -means که مراکز داده در آن توسط فرآیند بالا به صورت دستی داده شده است (نه به صورت تصادفی)، داده‌ها را خوش‌بندی کرد. شبیه کد شکل زیر روند فرآیند بالا را نشان می‌دهد که در آن ابتدا مقادیر ثابت‌ها (r_x) و مجموعه داده به عنوان ورودی گرفته می‌شود و پس از ساخت مراکز داده مطابق با تعاریف بالا، این مراکز برای خوش‌بندی در الگوریتم K -means استفاده می‌شود [79].

Inputs Dataset, Constants

Output Clusters

Steps

1. Initialize constants and density values
2. Make a new cluster center.
3. Update density values
4. If the sufficient number of clusters are not obtained, go to 2.
3. Clustering the dataset by k-means, using fix centers.

شکل ۱۱-۲. خوش‌بندی کاهشی

۱۱-۲-۵. الگوریتم خوش‌بندی Median K-Flat

الگوریتم Median K-Flat یا به اختصار MKF مجموعه داده‌ی $X = \{x_1, x_2, \dots, x_N\} \subseteq \mathbb{R}^D$ را به K خوش‌بندی X_1, X_2, \dots, X_K افزایش می‌کند که هر خوش‌بندی یک شبیه فضای d^{58} -بعدی تقریباً خطی می‌باشد. پارامتر P_i با فرض $1 \leq i \leq K$ ماتریسی با ابعاد $d \times D$ می‌باشد، که هر یک از خانه‌های آن تخمین شبیه فضای خطی متعامد^{۵۹} X_i می‌باشد. قابل به ذکر است که $P_i P_i^T = I_{d \times d}$ می‌باشد. در اینجا تخمین شبیه فضای خوش‌بندی X_1, X_2, \dots, X_K را با P_1, P_2, \dots, P_K نام‌گذاری می‌کنیم. مطابق تعاریف بالاتابع انرژی برای افزایش‌های $\{X_i\}_{i=1}^K$ بر اساس شبیه فضای $\{P_i\}_{i=1}^K$ به شکل زیر تعریف می‌شود [77].

$$\mathcal{E}\left(\{X_i\}_{i=1}^K, \{P_i\}_{i=1}^K\right) = \sum_{i=1}^K \sum_{x \in X_i} \|x - P_i^T P_i x\| \quad (13-2)$$

⁵⁸ Subspace

⁵⁹ Orthogonal

این الگوریتم سعی می‌کند تا مجموعه داده را به خوش‌های $\{X_i\}_{i=1}^K$ تبدیل کند به نحوی که تابع انرژی کمینه باشد. تا وقتی که سطوح تحت اساسی^{۶۰} به شکل شبیه فضای خطی هستند ما می‌توانیم به صورت فرضی المان‌های X را در یک حوضه واحد نرمال کنیم به طوری که $\|X_j\|=1$ برای $N \leq j \leq 1$ و تابع انرژی را به شکل زیر بیان کنیم: [77]

$$\mathcal{E}\left(\{X_i\}_{i=1}^K, \{P_i\}_{i=1}^K\right) = \sum_{i=1}^K \sum_{x \in X_i} \sqrt{\|x - P_i^T P_i x\|^2} = \sum_{i=1}^K \sum_{x \in X_i} \sqrt{1 - \|P_i x\|^2} \quad (14-2)$$

این الگوریتم برای کمینه‌سازی تابع انرژی الگوریتم MKF از روش کاهش گرادیان تصادفی استفاده می‌کند. مشتق تابع انرژی بر اساس ماتریس P_i به شرح زیر است:

$$\frac{\partial \mathcal{E}}{\partial P_i} = - \sum_{x \in X_i} \frac{P_i x x^T}{\sqrt{1 - \|P_i x\|^2}} \quad (15-2)$$

این الگوریتم نیاز به تطبیق P_i بر اساس مؤلفه‌ی متعامد مشتق P_i دارد. بخشی از مشتق که با شبیه فضای P_i موازی است به شرح زیر می‌باشد.

$$\frac{\partial \mathcal{E}}{\partial P_i} P_i^T P_i = - \sum_{x \in X_i} \frac{P_i x x^T P_i^T P_i}{\sqrt{1 - \|P_i x\|^2}} \quad (16-2)$$

از این روی مؤلفه متعامد برابر است با رابطه ۱۷-۲ می‌باشد.

$$dP_i = \sum_{x \in X_i} d_x P_i \quad (17-2)$$

در رابطه بالا $d_x P_i$ برابر با رابطه ۱۸-۲ است.

$$d_x P_i = - \frac{(P_i x x^T - P_i x x^T P_i^T P_i)}{\sqrt{1 - \|P_i x\|^2}} \quad (18-2)$$

با در نظر گرفتن محاسبات بالا، الگوریتم MKF تصمیم می‌گیرد که داده تصادفی X^* از مجموعه داده، عضو کدام P_{i^*} باشد، و از این طریق شروع به چیدن داده‌ها می‌کند. آن گاه، الگوریتم تابع P_{i^*} را به روز کند که در آن dt (مرحله زمانی) پارامتری است که توسط کاربر تعیین می‌شود. این فرآیند آن قدر تکرار می‌شود تا ضابطه همگرایی دیده شود. آنگاه هر نقطه از

⁶⁰ Underlying flats

مجموعه داده به نزدیکترین شبیه فضای $\{P_i\}_{i=1}^K$ که تعیین‌کننده خوش‌هاست اختصاص داده می‌شود.
شبیه کد زیر فرآیند الگوریتم MKF را نشان می‌دهد [77].

Input:

$X = \{x_1, x_2, \dots, x_N\} \subseteq \Re^D$: Data, normalized onto the unit sphere, d: dimension of subspaces K:

number of subspaces, $\{P_i\}_{i=1}^K$ the initialized subspaces. dt : step parameter.

Output: A partition of X into K disjoint clusters $\{X_i\}_{i=1}^K$

Steps:

1. Pick a random point x^* in X
2. Find its closest subspace P_{i^*} , where $i^* = \arg \max_{1 \leq i \leq K} \|P_i x\|$
3. Compute $d_{x^*} P_{i^*}$ by $d_{x^*} P_{i^*} = -\frac{(P_i x x^T - P_i x x^T P_i^T P_i)}{\sqrt{1 - \|P_i x\|^2}}$
4. Update $P_{i^*} \mapsto P_{i^*} - dt d_{x^*} P_{i^*}$
5. Orthogonalize P_{i^*}
6. Repeat steps 1-5 until convergence
7. Assign each x_i to the nearest subspace

شکل ۱۲-۲. شبیه کد الگوریتم MKF [77]

۶-۲-۱-۲-۲-۲-۶. الگوریتم خوش‌بندی مخلوط گوسی

یک مخلوط گوسی^{۶۱} یا همان GM را می‌توان ترکیب محدودی^{۶۲} از چگالی‌های گوسی دانست.
یک چگالی گوسی در فضای d-بعدی به ازای میانگین $m \in \Re^n$, توسط ماتریس هم‌واردایی^{۶۳} C با
ابعاد $d \times d$ به صورت زیر تعریف می‌شود: [83]

$$\phi(x; \theta) = (2\pi)^{-d/2} \det(C)^{-1/2} \exp(-(x-m)^T C^{-1}(x-m)/2) \quad (19-2)$$

⁶¹ Gaussian Mixture

⁶² Convex

⁶³ Covariance

در رابطه بالا θ پارامترهای m و C را تعریف می‌کند. از این روی k مؤلفه GM به صورت زیر تعریف می‌شود:

$$f_k(x) = \sum_{j=1}^k \pi_j \phi(x; \theta_j), \quad \text{with} \quad \sum_{j=1}^k \pi_j = 1 \quad \text{and} \quad \text{for } j \in \{1, \dots, k\}: \pi_j \geq 0 \quad (20-2)$$

در رابطه (20-2) پارامتر π وزن مخلوط کردن⁶⁴ و $(\phi(x; \theta_j))$ مؤلفه مخلوط می‌باشد. از آنجا که در مقایسه با تخمین چگالی غیر پارامتری، تعداد کمتری از توابع چگالی در تخمین چگالی مخلوط باید ارزیابی شود، از این روی ارزیابی چگالی کارآمدتر خواهد بود. علاوه بر آن، استفاده از اجرای محدودیت هموار کردن⁶⁵ بر روی برخی از مؤلفه‌های مخلوط در نتیجه‌ی چگالی به ما اجازه می‌دهد تا چگالی مستحکم‌تری را تخمین بزنیم. الگوریتم حداقل-انتظار⁶⁶ یا همان EM به ما اجازه به روز کردن پارامترهای K مؤلفه‌ی مخلوط را مطابق با مجموعه داده $X_n = \{x_1, \dots, x_n\} \in \mathbb{R}^d$ به ازای هر $x_i \in \mathbb{R}^d$ می‌دهد، به طوری که احتمال $P(j | x_i)$ هرگز کوچک‌تر از مخلوط جدید نشود. به روز کردن الگوریتم می‌تواند در یک فرآیند تکراری برای تمامی مؤلفه‌های $\{1, \dots, k\}$ مطابق با رابطه‌های زیر انجام شود: [83]

$$P(j | x_i) = \pi_j \phi(x_i; \theta) / f_k(x_i) \quad (21-2)$$

$$\pi_j = \sum_{i=1}^n P(j | x_i) / n \quad (22-2)$$

$$m_j = \sum_{i=1}^n P(j | x_i) x_i / (n \pi_j) \quad (23-2)$$

$$C_j = \sum_{i=1}^n P(j | x_i) (x_i - m_j) (x_i - m_j)^T / (n \pi_j) \quad (24-2)$$

در این تحقیق از روش پیشنهادی بومن و همکاران⁶⁷ برای پیاده‌سازی الگوریتم مخلوط گوسی استفاده شده است. از آنجایی که روش پیاده‌سازی و توضیحات مربوط به الگوریتم مخلوط گوسی در روش ترکیب مبتنی بر مخلوط استفاده می‌شود از این روی در بخش روش‌های ترکیب نتایج با تابع توافقی آن را بررسی خواهیم کرد.

⁶⁴ Mixing weight

⁶⁵ Smoothness constraint

⁶⁶ Expectation-Maximization

⁶⁷ <http://www.ece.purdue.edu/~bouman>

۲-۲-۲. معیارهای ارزیابی

در یادگیری با ناظر^{۶۸} ارزیابی راحت تر از یادگیری بدون ناظر است. برای مثال آن چیز که ما در رده‌بندی^{۶۹} باید ارزیابی کنیم مدلی است که ما توسط داده‌های^{۷۰} یادگیری به الگوریتم هوش مصنوعی^{۷۱} آموزش^{۷۲} داده‌ایم. در روش‌های با ناظر ورودی و خروجی داده معلوم است و ما بخشی از کل داده را برای آزمون جدا کرده و بخش دیگر را به عنوان داده یادگیری استفاده می‌کنیم و پس از تولید مدل مطلوب ورودی داده آزمون^{۷۳} را در مدل وارد کرده و خروجی مدل را با خروجی واقعی می‌سنجمیم^{۷۴}. از این روی معیارهای بسیاری برای ارزیابی روش‌های با ناظر ارائه شده‌اند.

در یادگیری بدون ناظر روش متفاوت است. در این روش هیچ شاخص معینی در داده جهت ارزیابی وجود ندارد و ما به دنبال دسته‌بندی کردن داده‌ها بر اساس شباهت‌ها و تفاوت‌ها هستیم. از این روی برخلاف تلاش‌های خیلی از محققان، ارزیابی خوشه‌بندی خیلی توسعه داده نشده است و به عنوان بخشی از تحلیل خوشه‌بندی رایج نشده است. در واقع، ارزیابی خوشه‌بندی یکی از سخت‌ترین بخش‌های تحلیل خوشه‌بندی است [33]. معیارهای عددی، یا شاخص‌هایی که برای قضایت جنبه‌های مختلف اعتبار یک خوشه به کار می‌روند، به سه دسته کلی تقسیم می‌شوند:

۱- شاخص خارجی^{۷۵} که مشخص می‌کند که کدام خوشه‌های پیداشده به وسیله الگوریتم خوشه‌بندی با ساختارهای خارجی تطبیق دارند. در این روش نیاز به اطلاعات اضافی مثل برچسب نقاط داده، داریم. آنتروپی یک مثالی از شاخص خارجی است.

۲- شاخص داخلی^{۷۶} که برای اندازه‌گیری میزان خوبی^{۷۷} یک ساختار خوشه‌بندی بدون توجه به اطلاعات خارجی به کار می‌رود. $SSE^{۷۸}$ یک نمونه از شاخص داخلی است.

⁶⁸ Supervised Learning

⁶⁹ Classification

⁷⁰ Dataset

⁷¹ Artificial intelligent algorithms

⁷² Learning

⁷³ Test Dataset

⁷⁴ Evaluation

⁷⁵ External Index

⁷⁶ Internal Index

⁷⁷ Goodness

⁷⁸ Sum of Squared Error

۳- شاخص نسبی^{۷۹} که برای مقایسه دو خوشبندی مختلف یا دو خوشه مختلف به کار می‌رود. اغلب یک شاخص خارجی یا داخلی برای این تابع استفاده می‌شود. برای مثال، دو خوشبندی K -means می‌توانند با مقایسه SSE یا آنتروپی شان مقایسه شوند.

این فصل تعدادی از مهم‌ترین و رایج‌ترین روش‌های به کار رفته برای ارزیابی خوشبندی را مرور خواهد کرد.

۱-۲-۲-۲. معیار SSE

یک معیار داخلی ارزیابی خوشبندی، مثل SSE ، می‌تواند برای ارزیابی یک خوشبندی نسبت به خوشبندی دیگر به کار رود. به علاوه، یک معیار داخلی اغلب می‌تواند برای ارزیابی یک خوشبندی کامل یا یک خوشه تنها به استفاده شود. این اغلب به خاطر این است که این روش، سعی می‌کند تا میزان خوبی کلی خوشبندی را به عنوان یک جمع وزن‌دار از خوبی‌های هر خوشه در نظر می‌گیرد. SSE با استفاده از رابطه ۲۵-۲ محاسبه می‌شود [68].

$$SSE = \sum_{i=1}^K \sum_{x \in c_i} \sum_{j=1}^n (m_j^i - x_j)^2 \quad (25-2)$$

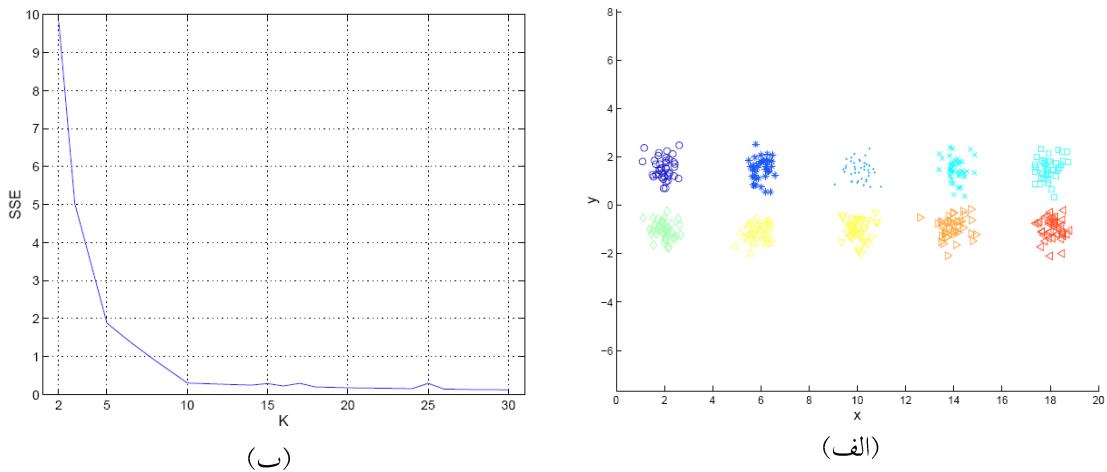
که λ یک نقطه داده در خوشه C_i است و x_j ، j -امین ویژگی از داده X است. m_j^i ، j -امین ویژگی از مرکز خوشه C_i می‌باشد. برای مقایسه دو خوشبندی مختلف روی یک داده با یک تعداد مشابه، تنها مقایسه مقدارهای متناظر SSE آن‌ها کافی است. هر چه مقدار SSE کمتر باشد، آن خوشبندی بهتر خواهد بود. البته، وقتی تعداد نقاط داده در دو خوشه متفاوت باشند، مقایسه مستقیم از روی مقدار SSE خوب نخواهد بود. بنابراین، $mean SSE$ یک خوشه معیار مناسب تری برای مقایسه است. رابطه ۲۶-۲ این معیار را نشان می‌دهد که در آن مقدار N تعداد کل نمونه‌هاست [68].

$$mean SSE = \frac{SSE}{N} \quad (26-2)$$

تعداد درست خوشه‌ها در الگوریتم K -means، اغلب می‌تواند با استفاده از نگاه کردن به منحنی SSE مشخص شود. این منحنی با رسم مقادیر SSE به ازای K ‌های مختلف به دست می‌آید. تعداد

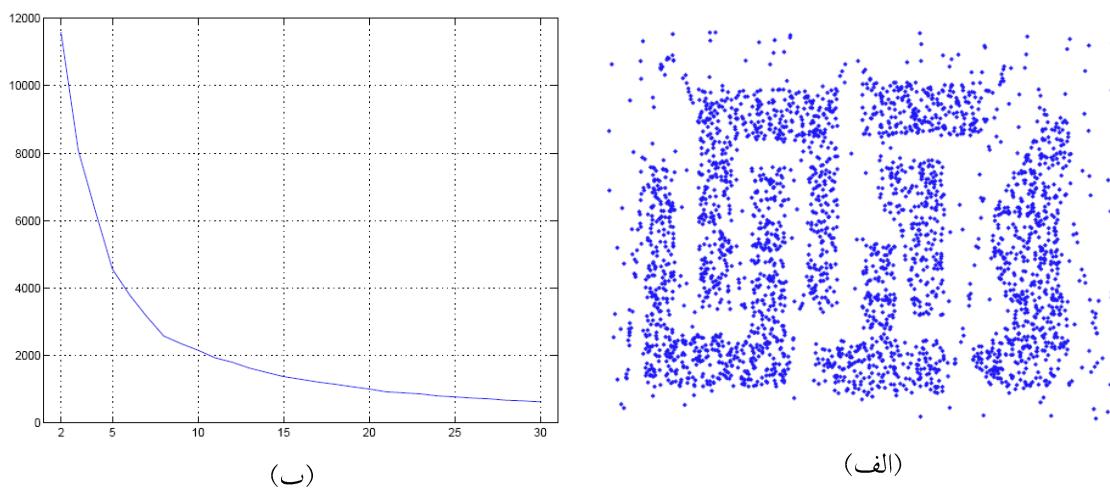
⁷⁹ Relative Index

خوشه‌های بهینه با توجه به منحنی SSE ، قابل K ای است که به ازای آن نرخ کاهش مقدار SSE ، چشم‌پوشی شود. شکل ۱۳-۲-ب منحنی SSE را برای داده‌های شکل ۱۳-۲-الف، نشان می‌دهد.



شکل ۱۳-۲. (الف) مجموعه داده با تعداد ۱۰ خوشه واقعی. (ب) منحنی SSE مربوطه [68]

همان طور که از شکل ۱۳-۲-ب برمی‌آید، برای مقادیر k ‌های از صفر تا ۱۰ شیب منحنی نسبت به بقیه مقادیر k ، تندتر می‌باشد. این امر نشان‌دهنده آن است که مقدار $k = 10$ یک مقدار بهینه برای تعداد خوشه‌ها می‌باشد.



شکل ۱۴-۲. (الف) مجموعه داده (ب) منحنی SSE مربوطه [2]

شکل ۱۴-۲-ب نیز منحنی SSE را برای داده‌های شکل ۱۴-۲-الف، نشان می‌دهد. مشاهده می‌شود که در این داده‌ها، چون تعداد خوشه‌ها نسبت به شکل ۱۴-۲-الف کاملاً گویا نیست، بنابراین، منحنی SSE آن نیز نرم تر خواهد بود. اما با توجه به شکل ۱۴-۲-ب، می‌توان گفت که تعداد $k = 8$ نسبتاً خوب باشد. چون منحنی برای k ‌های بعد از ۸، دارای شیب کندتری خواهد شد. با توجه به نتایج

فوق می توان گفت که اگرچه منحنی SSE برای همه مسایل نمی تواند جواب بهینه برای تعداد k بدهد، اما می تواند به عنوان یک معیار خوب برای این امر مطرح باشد.

۲-۲-۲-۲. معیار اطلاعات متقابل نرمال شده

معیار اطلاعات متقابل (MI^*)^{۸۰} توسط کاور و توماس [71] معرفی شد که یک روش جهت اندازه گیری کیفیت اطلاعات آماری مشترک بین دو توزیع است. از آنجایی که این معیار وابسته به اندازه خوشها است در [54] روشی جهت نرمال سازی آن ارائه شده است. فرد و جین [19] روش نرمال سازی اطلاعات متقابل را اصلاح کردند و آن را تحت عنوان اطلاعات متقابل نرمال (NMI^*)^{۸۱} ارائه داده اند. رابطه ۲۷-۲ اطلاعات متقابل نرمال شده را نشان می دهد [1, 2, 19].

$$NMI(P_1, P_2) = \frac{MI(P_1, P_2)}{\frac{-1}{2m} \left(\sum_{i=0}^1 p_{i.} \log \frac{p_{i.}}{m} + \sum_{j=0}^1 p_{.j} \log \frac{p_{.j}}{m} \right)} \quad (27-2)$$

در رابطه ۲۷-۲ پارامتر m کل نمونه ها است و $p_{i.}$ یعنی افزایه ای که اندیس آنها شامل i با تمام مقادیر زمی باشد و $p_{.j}$ یعنی افزایه ای که تمام مقادیر j با و اندیس j را شامل شود. $MI(P_1, P_2)$ از رابطه ۲۸-۲ محاسبه می شود [1, 2, 19].

$$MI(P_1, P_2) = \sum_{i=0}^1 \sum_{j=0}^1 \frac{r_{ij}}{m^2} \log \frac{mp_{ij}}{r_{ij}} \quad (28-2)$$

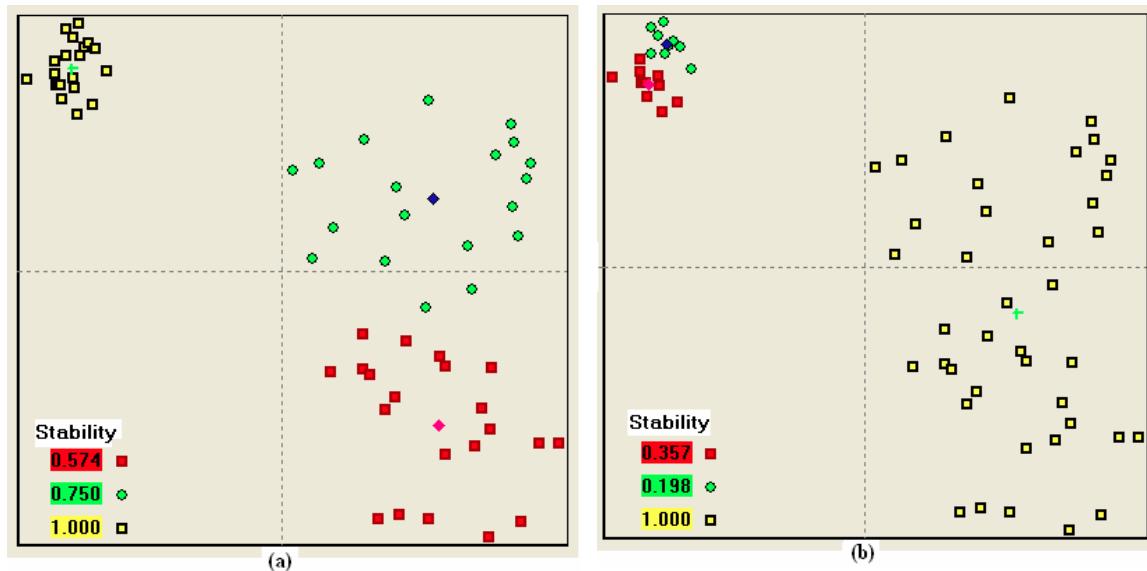
$$r_{ij} = p_{i.}p_{.j} \quad , \quad p_{i.} = p_{i0} + p_{il} \quad , \quad p_{.j} = p_{0j} + p_{1j}$$

در صورتی که دو افزای به صورت $P_2 = \{C_2, D/C_2\}$ و $P_1 = \{C_1, D/C_1\}$ که در آن D کل داده و C_i خوش اول و D/C_i خوش دوم هر یک از افزایها باشد آنگاه p_{11} نشان دهنده تعداد نمونه های مشترک موجود در C_1 و C_2 می باشد، p_{10} نشان دهنده تعداد نمونه های مشترک موجود در C_1 و D/C_2 می باشد و p_{00} نشان دهنده تعداد نمونه های مشترک موجود در D/C_1 و D/C_2 می باشد. در واقع $p_{i.}$ و $p_{.j}$ به ترتیب بیانگر کل نمونه های موجود در C_1 و C_2 می باشد [1].

⁸⁰ Mutual Information

⁸¹ Normalized MI

شکل ۱۵-۲ دو افزار اولیه را نشان می‌دهد که میزان پایداری برای هر کدام از خوش‌های به دست آمده هم محاسبه شده است. در این مثال الگوریتم K -means به عنوان الگوریتم خوش‌بندی اولیه انتخاب شده است و تعداد خوش‌های اولیه برابر با سه نیز به عنوان پارامتر آن از قبل مشخص شده است. همچنین، در این مثال تعداد افزارهای موجود در مجموعه مرجع برابر با ۴۰ می‌باشد. در ۳۶ افزار نتایجی مشابه با شکل ۱۵-۲ (a) و در ۴ حالت باقیمانده نیز نتایجی مشابه با شکل ۱۵-۲ (b) حاصل شده است [1].



شکل ۱۵-۲. دو افزار اولیه با تعداد سه خوش. (a) خوش‌بندی درست (b) خوش‌بندی نادرست [1]

از آن جایی که در مجموعه مرجع در ۹۰٪ موقع، داده‌های متراکم گوش بالا-چپ از شکل ۱۵-۲ در یک خوش مجزا گروه‌بندی شده‌اند، بنابراین این خوش باید مقدار پایداری بالایی را به خود اختصاص دهد. اگرچه این مقدار باید دقیقاً برابر با یک باشد (چون در همه موارد این خوش درست تشخیص داده نشده است)، مقدار پایداری با روش متداول اطلاعات متقابل نرمال شده مقدار یک را بر می‌گرداند. از آن جایی که ادغام دو خوش سمت راست تنها در ۱۰٪ موارد مانند شکل ۱۵-۲ (b) اتفاق افتاده است، خوش حاصل باید مقدار پایداری کمی به دست آورد. اگر چه خوش حاصل از ادغام دو خوش سمت راستی، به ندرت (۱۰٪ موارد) در مجموعه مرجع دیده شده است، مقدار پایداری برای این خوش نیز برابر با یک به دست می‌آید. در اینجا مشکل روش متداول محاسبه پایداری با استفاده از اطلاعات متقابل نرمال شده ظاهر می‌شود. از آنجایی که معیار اطلاعات متقابل نرمال شده یک معیار مترکن است، مقدار پایداری خوش بزرگ ادغامی سمت راست (با ۱۰٪ تکرار) دقیقاً برابر با میزان پایداری خوش متراکم گوش بالا-چپ (با ۹۰٪ تکرار) به دست می‌آید. به عبارت دیگر در مواردی که داده‌های دو خوش مکمل یکدیگر باشند، یعنی اجتماع داده‌های آنها شامل کل

مجموعه داده شود و اشتراک داده‌های آنها نیز تهی باشد، مقدار پایداری برای هر دو به یک اندازه برابر به دست می‌آید. از دیدگاه دیگر، این اتفاق زمانی رخ می‌دهد که تعداد خوش‌های تشکیل‌دهنده مجموعه C_2 در خوش‌بندی مرجع عددی بیشتر از یک باشد. هر زمان که C_2 با ادغام دو یا بیشتر از خوش‌های به دست آید، منجر به نتایج نادرست در مقدار پایداری می‌شود. ما این مشکل را تحت عنوان مشکل تقارن در اطلاعات متقابل نرم‌الشده می‌شناسیم. در سال‌های اخیر روش‌هایی جهت حل این مشکل ارائه شده‌اند که یکی از آنها را علیزاده و همکاران در [9, 1] ارائه داده‌اند که در آن بزرگ‌ترین خوش‌های بین مجموعه مرجع (که بیش از نصف نمونه‌هایش در خوش‌هایش مورد مقایسه وجود دارد) جایگزین اجتماع همه خوش‌های می‌شود که ما آن را با عنوان روش Max می‌شناسیم. روش دیگر جهت رفع این مشکل معیار APMM^{۸۲} می‌باشد. در ادامه به بررسی این معیار می‌پردازیم [1, 8, 67].

۳-۲-۲-۲. معیار APMM

برخلاف معیار NMI که برای اندازه‌گیری شباهت دو افزار طراحی شده است معیار APMM روشی برای اندازه‌گیری میزان شباهت یک خوش‌های در یک افزار است که توسط علیزاده و همکاران [8, 67] معرفی شده است رابطه ۲۹-۲ این معیار را معرفی می‌کند.

$$APMM(C_i^a, P^{b*}) = \frac{-2 \log \left(\frac{n}{n_i^a} \right) \sum_{j=1}^{k_b^*} n_j^{b*}}{n_i^a \log \left(\frac{n_i^a}{n} \right) + \sum_{j=1}^{k_b^*} n_j^{b*} \log \left(\frac{n_j^{b*}}{n} \right)} \quad (29-2)$$

در رابطه ۲۹-۲ پارامتر C_i^a خوش‌های i در افزار P^a می‌باشد و P^{b*} افزار متناظر با خوش‌های C_i در خوش‌بندی P^b است. پارامتر n تعداد کل نمونه‌های مجموعه داده و n_{ij}^{ab} تعداد نمونه‌های مشترک بین خوش‌های $C_i^b \in P^b$ و $C_i^a \in P^a$ می‌باشد. همچنین، k_b^* تعداد خوش‌های موجود در افزار P^{b*} می‌باشد. در این روش برای محاسبه پایداری خوش‌های C_i از رابطه ۳۰-۲ استفاده می‌کنیم [8, 67].

$$AAPMM(C_i) = \frac{1}{M} \sum_{j=1}^M APMM(C_i^a, P_j^{b*}) \quad (30-2)$$

⁸² Alizadeh-Parvin-Moshki-Minaei

در رابطه ۲-۳۰ پارامتر $P_j^{b^*}$ نشان‌دهنده زیرا افزار از مجموعه مرجع است و M تعداد کل افزارها است [67, 8]. از آنجایی که این معیار برای ارزیابی شباهت یک خوشه است می‌توان هم برای ارزیابی خوشه و هم برای ارزیابی افزار استفاده کرد. جهت استفاده از این معیار برای ارزیابی یک افزار کافی است آن را برای تک‌تک خوشه‌های آن افزار استفاده کنیم و در نهایت از کل مقادیر میانگین بگیریم.

۲-۳. خوشبندی ترکیبی

کلمه 'Ensemble'، ریشه فرانسوی دارد و به معنی باهم بودن یا در یک زمان می‌باشد و معمولاً اشاره به واحدها و یا گروه‌های مکملی دارد که باهم در اجرای یک کار واحد همکاری می‌کنند. ترکیب تاریخ طولانی در دنیای واقعی دارد، نظریه هیئت‌منصفه‌ی کندورست که در سال ۱۷۸۵ میلادی مطرح شده است و این ایده را مطرح می‌کند که، احتمال نسبی درستی نظر گروهی از افراد (رأی اکثریت) بیشتر از نظر هر یک از افراد به تنها یی می‌باشد را می‌توان دلیلی برای ترکیب نتایج در دنیای واقعی دانست [27, 10]. خوشبندی ترکیبی روشنی جدید در خوشبندی می‌باشد که از ترکیب نتایج روش‌های خوشبندی متفاوت به دست می‌آید از آنجایی که اکثر روش‌های خوشبندی پایه روی جنبه‌های خاصی از داده‌ها تاکید می‌کنند، در نتیجه روی مجموعه داده‌های خاصی کارآمد می‌باشند. به همین دلیل، نیازمند روش‌هایی هستیم که بتواند با استفاده از ترکیب این الگوریتم‌ها و گرفتن نقاط قوت هر یک، نتایج بهینه‌تری را تولید کند. هدف اصلی خوشبندی ترکیبی جستجوی نتایج بهتر و مستحکم‌تر، با استفاده از ترکیب اطلاعات و نتایج حاصل از چندین خوشبندی اولیه است [18, 54]. خوشبندی ترکیبی می‌تواند جواب‌های بهتری از نظر استحکام^{۸۳}، نو بودن^{۸۴}، پایداری^{۸۵} و انعطاف‌پذیری^{۸۶} نسبت به روش‌های پایه ارائه دهد [3, 21, 54, 57]. به طور خلاصه خوشبندی ترکیبی شامل دو مرحله اصلی زیر می‌باشد: [34, 54]

۱- تولید نتایج متفاوت از خوشبندی‌ها، به عنوان نتایج خوشبندی اولیه بر اساس اعمال روش‌های مختلف که این مرحله را، مرحله ایجاد تنوع یا پراکندگی^{۸۷} می‌نامند.

⁸³ Robustness

⁸⁴ Novelty

⁸⁵ Stability

⁸⁶ Flexibility

⁸⁷ Diversity

۲- ترکیب نتایج به دست آمده از خوشبندی‌های متفاوت اولیه برای تولید خوش نهایی؛ که این کار توسط تابع توافقی^{۸۸} (الگوریتم ترکیب‌کننده) انجام می‌شود.

۱-۳-۲. ایجاد تنوع در خوشبندی ترکیبی

در خوشبندی ترکیبی، هرچه خوشبندی‌های اولیه نتایج متفاوت تری ارائه دهنده نتیجه نهایی بهتری حاصل می‌شود. در واقع هرچه داده‌ها از جنبه‌های متفاوت‌تری مطالعه و بررسی شوند (تشخیص الگوهای پنهان داده) نتیجه نهایی که از ترکیب این نتایج حاصل می‌شود متعاقباً دارای دقت بالاتری خواهد بود که این امر منجر به کشف دانش ضمنی پنهان در داده نیز خواهد شد. تنوع در این بخش به این معنا می‌باشد که با استفاده از روش‌های متفاوت مجموعه داده را از دیدگاه‌های گوناگونی مورد بررسی قرار دهیم. در این فصل برای ایجاد پراکندگی در بین نتایج حاصل چند راهکار مختلف پیشنهاد می‌کنیم و به بررسی مطالعات انجام‌شده در هر یک از آن‌ها می‌پردازیم. راه‌های مختلفی برای ایجاد پراکندگی در خوشبندی ترکیبی وجود دارد که عبارت‌اند از:

۱- استفاده از الگوریتم‌های متفاوت خوشبندی.

۲- تغییر مقادیر اولیه و یا سایر پارامترهای الگوریتم خوشبندی انتخاب شده.

۳- انتخاب بعضی از ویژگی‌های داده‌ها یا ایجاد ویژگی‌های جدید.

۴- تقسیم‌بندی داده‌های اصلی به زیرمجموعه‌هایی متفاوت و مجزا.

در حقیقت به خاطر ماهیت بدون ناظر بودن مسئله خوشبندی این اصل که آیا پراکندگی به وجود آمده مفید می‌باشد یا مفید نیست را نمی‌تواند مورد مطالعه قرارداد اما نتایج تجربی نشان داده است که ایجاد پراکندگی در خوشبندی‌های اولیه به طور معمول موجب بهبود خوشبندی در اکثر موقعیت‌ها می‌شود لذا در روش‌های ارائه‌شده هدف تنها بررسی مجموعه داده از زوایای مختلف است [42].

⁸⁸ Consensus Function

۱-۱-۳-۲. استفاده از الگوریتم‌های مختلف خوشبندی ترکیبی

به طور معمول بیشتر روش‌های خوشبندی ترکیبی از الگوریتم K -means جهت خوشبندی اولیه خود استفاده می‌کنند [37, 47, 56, 57]. اما در روش‌های ارائه شده نشان داده شده است که با توجه به رفتار هر مجموعه داده گاهی اوقات یک روش خوشبندی خاص پیدا می‌شود که دقیق‌تر از K -means برای بعضی از مجموعه داده‌ها می‌دهد [54]. اما الگوریتم K -means به دلیل سادگی و توانایی مناسب در خوشبندی همواره به عنوان انتخاب اول در خوشبندی ترکیبی مورد مطالعه قرار گرفته است. نکته مهمی که در انتخاب الگوریتم‌ها باید به آن دقیق‌تر این است که الگوریتم‌هایی همانند K -means که بر اساس فاصله اقلیدسی تمامی ویژگی‌ها کار می‌کنند، در صورتی که حتی یک ویژگی یک نمونه دارای یک مقدار غیرمنتظره باشد، نمونه به طور نادرست دسته‌بندی می‌شود. با توجه به این مسئله می‌توان از روش‌هایی مشابه این الگوریتم‌ها که مقاوم در برابر نویز هستند جهت رسیدن به پایداری و کیفیت بیشتر استفاده کرد. نکته دیگری که در انتخاب الگوریتم‌های پایه باید به آن توجه کرد این است که برخی از روش‌ها همانند الگوریتم‌های سلسله مراتبی پیوندی^{۸۹} همواره با تکرار مکرر روی یک داده یک جواب منحصر به فرد ایجاد می‌کنند که در صورت ایجاد نتایج با این گونه الگوریتم‌ها باید فقط یکی از هر نوع آن را در ساخت نتایج نهایی استفاده کرد.

۲-۱-۳-۲. تغییر پارامترهای اولیه خوشبندی ترکیبی

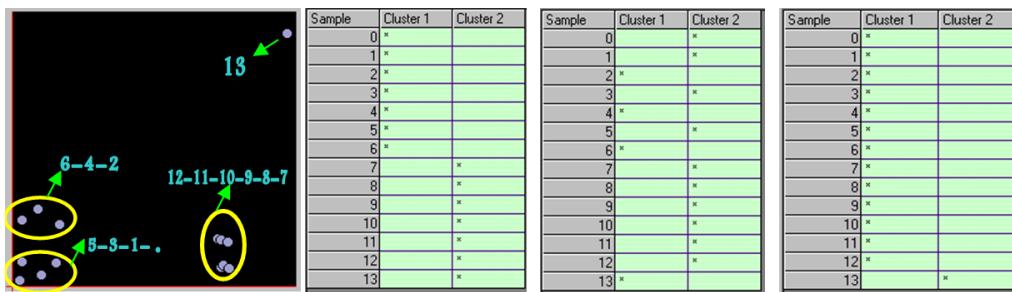
یکی دیگر از راه‌های افزایش پراکندگی تغییر پارامترهای اولیه الگوریتم‌های خوشبندی می‌باشد. برای مثال در الگوریتم K -means می‌توان با تغییر تعداد خوشبندی‌ها در الگوریتم، یا تعداد دفعات تکرار^{۹۰} اجرای الگوریتم و یا تغییر نمونه‌های اولیه^{۹۱} الگوریتم میزان پراکندگی را افزایش داد. در شکل ۱۶-۲ اثر نمونه‌های اولیه در خوشبندی نهایی به وضوح قابل مشاهده می‌باشد. در شکل زیر در سمت چپ ابتدا نحوه توزیع نمونه‌ها^{۹۲} نمایش داده شده است و سپس نتایج سه بار اجرای مختلف الگوریتم با سه نمونه شروع مختلف نمایش داده شده است [2, 6].

⁸⁹ Linkage

⁹⁰ Iterative

⁹¹ Seed Points

⁹² Scatter Plot



شکل ۱۶-۲. نمونه‌های اولیه در نتایج الگوریتم K -means. شکل‌ها به ترتیب از چپ به راست ۱) نمایش فضایی ۱۴ نمونه پراکنده در فضای ۲) نتایج به دست آمده با نمونه‌های اولیه ۱ و ۸ ۳) نتایج به دست آمده با نمونه‌های اولیه ۲ و ۳ ۴) نتایج به دست آمده با نمونه‌های اولیه ۱ و ۱۳ [2].

۳-۱-۳-۲. انتخاب یا تولید ویژگی‌های جدید

استفاده از برخی از ویژگی‌های کل فضای مجموعه داده و یا تولید ویژگی‌های جدید یکی دیگر از راهکارهای افزایش پراکندگی در خوشبندی ترکیبی می‌باشد. بسیاری از مطالعات در حیطه طبقه‌بندی اطلاعات اقدام به انتخاب زیرمجموعه‌ای از ویژگی‌ها می‌نماید که باعث افزایش میزان پراکندگی، کاهش حجم محاسبات و بالا بردن دقیق‌تری طبقه‌بندی کننده می‌شود [54]. ولی به دلیل ماهیت بدون ناظر بودن مسئله در خوشبندی، انتخاب زیرمجموعه‌ای از ویژگی‌ها کمتر مورد توجه بوده است و بیشتر سعی در تولید ویژگی‌های جدید بوده است. روش‌های گوناگونی برای تولید ویژگی و استفاده از آن در خوشبندی ترکیبی وجود دارد که ساده‌ترین آن‌ها نرم‌افزاری داده‌ها می‌باشد. معمولاً داده‌های مسائلی که از فاصله اقلیدسی برای خوشبندی آن‌ها استفاده می‌شود نرم‌افزاری نیست. نتایج تجربی نشان داده است که علیرغم اینکه نرم‌افزاری داده‌ها در بعضی مواقع موجب بهبود کار می‌شود در بعضی موارد موجب افت کارایی یک روش می‌شود [12].

۴-۱-۳-۲. انتخاب زیرمجموعه‌ای از مجموعه داده اصلی

یکی از راههای به دست آوردن این پراکندگی استفاده از تعداد محدودی از نمونه‌ها به جای کل نمونه‌ها می‌باشد که این امر دو مزیت دارد اول کاهش میزان محاسبات و دوم افزایش پراکندگی. روش‌های متعددی تاکنون برای ایجاد زیرمجموعه‌ها پیشنهاد گردیده است. در روش‌های معمولی، شانس نمونه‌ها برای انتخاب شدن در زیرمجموعه برابر $(1/N)$ می‌باشد [57]. یکی از روش‌های

معروف در انتخاب زیرمجموعه‌ای از مجموعه داده اصلی نمونه‌برداری^{۹۳} می‌باشد که می‌تواند با جایگزینی یا بدون جایگزینی و یا با انتخاب تصادفی^{۹۴} باشد.

۲-۳-۲. ترکیب نتایج با تابع توافقی

ترکیب نتایج خوشبندی‌های اولیه (پایه) و دستیابی به نتیجه نهایی یکی از مهم‌ترین مراحل خوشبندی ترکیبی می‌باشد. روش‌های گوناگونی برای ترکیب نتایج خوشبندی‌های اولیه مختلف و ایجاد خوشبندی نهایی وجود دارد که در زیر به معرفی چند روش جدید و معروف در این زمینه می‌پردازیم ولی به طور کل می‌توان آنها را در سه گروه مبتنی بر ابر گراف‌ها، روش رأی‌گیری و روش‌های مبتنی بر ماتریس همبستگی دسته‌بندی کرد.

۱-۲-۳-۲. روش مبتنی بر مدل مخلوط

این روش توسط تاپچی و همکاران [57] معرفی شده است. فرض کنید یک دسته N تایی از نقاط داده $\{x_1, \dots, x_N\}$ و یک دسته H تایی افزای $\{\pi_1, \dots, \pi_H\} \Pi$ از اشیاء X داریم. افزایهای متفاوت از X برای هر نقطه از x_i یک مجموعه از برصسب‌ها را برمی‌گرداند:

$$x_i = \{\pi_1(x_i), \pi_2(x_i), \dots, \pi_H(x_i)\}, i = 1, \dots, N \quad (31-2)$$

در اینجا، H خوشبندی مختلف نشان داده شده است و $(x_i)_j \pi$ نشان‌دهنده برصسب تخصیص یافته x_i توسط الگوریتم j -ام است. روش مدل مخلوط^{۹۵}، با استفاده از تعداد محدودی از مدل‌های مخلوط شده با توجه به احتمال وقوع برصسب‌های خوشه $y = \pi(x)$ از الگو یا اشیاء x روشی برای حل تابع توافقی پیشنهاد می‌دهد. فرض اصلی در این روش این است که برصسب‌های π مدلی از ترسیم تغییرهای تصادفی^{۹۶} از یک توصیف توزیع احتمال^{۹۷} در یک مخلوط از مؤلفه‌های متراکم چند متغیره است.

⁹³ Sampling

⁹⁴ Random

⁹⁵ Mixture Model

⁹⁶ Random variables drawn

⁹⁷ Multivariate component densities

$$P(y_i | \Theta) = \sum_{m=1}^M \alpha_m P_m(y_i | \theta_m) \quad (32-2)$$

در رابطه ۳۲-۲ مؤلفه توسط پارامتر θ_m تعریف شده است. M مؤلفه در مخلوط با خوشه‌های افزای توافقی π شناسایی می‌شوند. ضریب مخلوط α_m متناظر با احتمالات قبلی از خوشه‌ها تعیین می‌شود. در این مدل نقاط داده $\{y_i\}$ بر اساس تولید دو مرحله ذیل فرض می‌شود: اول، برای طراحی مؤلفه بر اساس احتمال جرم تابع $P_m(y | \theta_m)$. تمام داده $Y = \{y_i\}_{i=1}^N$ به صورت مستقل و به صورت یکسان توزیع شده فرض می‌شوند. این امر اجازه می‌دهد تا برای تعیین پارامترهای $\{\alpha_1, \dots, \alpha_M, \theta_1, \dots, \theta_M\}$ در مجموعه داده Y از نمایش تابع احتمال لگاریتمی رابطه ۳۳-۲ استفاده کنیم.

$$\log L(\Theta | Y) = \log \prod_{i=1}^N P(y_i | \Theta) = \sum_{i=1}^N \log \sum_{m=1}^M \alpha_m P_m(y_i | \theta_m) \quad (33-2)$$

با این کار هدف خوشبندی توافقی به یک مسئله تخمین احتمال حداقل فرموله شده است. برای پیدا کردن بهترین چگالی مخلوط مناسب برای داده Y ، باید تابع احتمال نسبت به پارامتر ناشناس Θ بیشینه شود. رابطه ۳۴-۲ نشان‌دهنده این مسئله می‌باشد:

$$\Theta^* = \arg \max \log L(\Theta | Y) \quad (34-2)$$

مرحله مهم بعدی مشخص کردن تراکم مؤلفه‌شرطی $P_m(y | \theta_m)$ است. توجه داشته باشید، که مسئله اصلی در خوشبندی در فضای داده X با کمک الگوریتم‌های متعدد به فضای ویژگی‌های جدید چند متغیره (x, π) تبدیل شده است. برای ساده‌سازی بیشتر مسئله، یک استقلال مشروط برای ساخت مؤلفه‌های بردار y فرض می‌شود، برای مثال احتمال شرطی زیر را می‌توان برای y در نظر گرفت.

$$P_m(y_i | \theta_m) = \prod_{j=1}^H P_m^{(j)}(y_{ij} | \theta_m^{(j)}) \quad (35-2)$$

در توجیه این کار، می‌توان ذکر کرد که حتی اگر الگوریتم‌های خوشبندی متفاوت (که باز شاخص گذاری می‌شوند) واقعاً مستقل نباشند، تقریب به وجود آمده در (۲۱-۲) را می‌تواند با کارایی عالی در طبقه‌بندی بیز ساده در حوزه‌های گستته توجیه کرد [43]. هدف نهایی تخصیص برچسب‌های مجزا به داده X از طریق مسیریابی غیرمستقیم برآورد چگالی Y است. الگوهای تخصیص یافته به خوشه‌ها

در π حساسیت کمتری به تقریب استقلال شرطی که با مقادیر احتمال $P(y_i | \Theta)$ محاسبه می‌شود دارد. آخرین مرحله از مدل مخلوط انتخاب احتمال چگالی $P_m^{(j)}(y_{ij} | \theta_m^{(j)})$ برای مؤلفه‌های بردارهای y است. تا موقع‌هایی که متغیرهای y_{ij} دارای ارزش اسمی از یک دسته از برچسب‌های خوش در افزای π باشد، طبیعی است که آنها را به عنوان نتایج حاصل از یک آزمایش چندجمله‌ای زیر فرض کنیم:

$$P_m^{(j)}(y | \theta_m^{(j)}) = \prod_{k=1}^{K(j)} v_{jm}(k)^{\delta(y,k)} \quad (36-2)$$

در اینجا، بدون، فراموش کردن اصل کلی، برچسب‌های خوش‌ها در π توسط اعداد صحیح در $\{1, \dots, K(j)\}$ انتخاب می‌شود. برای وضوح بیشتر این مطلب، باید توجه داشته باشید که احتمال نتایج توسط $v_{jm}(k)$ تعریف می‌شود و نتایج حاصل شامل همه مقادیر ممکن از برچسب‌های y_{ij} در افزای π است. همچنین، خلاصه احتمالات به صورت زیر است:

$$\sum_{k=1}^{K(j)} v_{jm}(k) = 1, \forall j \in \{1, \dots, H\}, \forall m \in \{1, \dots, M\} \quad (37-2)$$

به عنوان مثال، اگر زامین افزای فقط شامل دو خوش باشد، و برچسب‌های ممکن ۰ و ۱ باشند، آنگاه رابطه (37-2) می‌تواند به صورت زیر ساده شود:

$$P_m^{(j)}(y | \theta_m^{(j)}) = v_{jm}^y (1 - v_{jm})^{1-y} \quad (38-2)$$

بیشینه کردن مسئله احتمال⁹⁸ در رابطه (38-2) عموماً موقع‌هایی که تمام پارامترهای $\Theta = \{\alpha_1, \dots, \alpha_M, \theta_1, \dots, \theta_M\}$ معلوم نباشند، نمی‌تواند با فرم بسته حل شود. لیکن، تابع احتمال رابطه (38-2) می‌تواند با به کارگیری الگوریتم EM بهینه شود. به منظور اتخاذ الگوریتم EM ، داده پنهان Z و احتمال کل داده‌های (Y, Z) فرض می‌شود. توزیع Z باید مطابق با مقادیر مشاهده شده Z باشد:

$$\log P(Y | \Theta) = \log \sum_{z \in Z} P(Y, z | \Theta) \quad (39-2)$$

⁹⁸ Likelihood

اگر مقدار z_i معلوم باشد آنگاه می‌توان فوراً گفت که M مؤلفه‌ی مخلوطی در تولید نقطه y_i استفاده شده است. این به این معنی است که به ازای هر نقطه y_i مشاهده شده، یک متغیر بردار پنهان $z_i = \{z_{i1}, z_{i2}, \dots, z_{iM}\}$ وجود دارد به طوری که y_i متعلق به m -امین مؤلفه باشد و در غیر این صورت y_i می‌باشد. برای نوشتمن احتمال داده کامل رابطه زیر مناسب است:

(۴۰-۲)

$$\log L(\Theta | Y, Z) = \log \prod_{i=1}^N P(y_i, z_i | \Theta) = \log \prod_{i=1}^N \prod_{m=1}^M (\alpha_m P_m(y_i | \theta_m))^{z_{im}} = \sum_{i=1}^N \sum_{m=1}^M z_{im} \log \alpha_m P_m(y_i | \theta_m)$$

مطابق با روش عمومی EM ، بایدتابع کمکی $Q(\Theta; \Theta')$ که به عنوان یک حد پایین از مشاهده احتمال داده در رابطه (۳۳-۲) به کار گرفته می‌شود تعریف کنیم:

$$Q(\Theta; \Theta') = \sum_z \log(P(Y, z | \Theta)) p(z | Y, \Theta') \quad (41-2)$$

در روش کلاسیک تحلیل همگرایی الگوریتم EM [45, 16]، رابطه بیشینه سازی تابع $Q(\Theta; \Theta')$ با توجه به Θ معادل افزایش تابع احتمال مشاهده شده در رابطه (۳۳-۲) است. ارزیابی $Q(\Theta; \Theta')$ اولین مرحله الگوریتم EM است. جایگزین رابطه (۴۰-۲) در تعریف Q برابر است با:

(۴۲-۲)

$$Q(\Theta, \Theta') = \sum_z \sum_{i=1}^N \sum_{m=1}^M z_{im} \log \alpha_m P_m(y_i | \theta_m) p(z | Y, \Theta') = \sum_{i=1}^N \sum_{m=1}^M E[z_{im}] \log \alpha_m P_m(y_i | \theta_m)$$

در رابطه بالا برای تخمین زدن پارامترهای Θ' مانیاز به تعریف $E[z_{im}]$ به صورت زیر داریم:

$$E[z_{im}] = \sum_z z_{im} p(z | Y, \Theta') = \frac{\alpha'_m P_m(y_i | \theta'_m)}{\sum_{n=1}^M \alpha'_n P_n(y_i | \theta'_n)} \quad (43-2)$$

اینجا، حدس اولیه در مورد پارامترهای $\Theta' = \{\alpha'_1, \dots, \alpha'_M, \theta'_1, \dots, \theta'_M\}$ برای محاسبه مقادیر مورد انتظار استفاده می‌شود. با توجه به نوع تراکم چگالی مؤلفه شرطی $P_m(y_i | \theta_m)$ در رابطه (۳۵-۲) و (۳۶-۲)، ما برای E -امین مرحله از الگوریتم رابطه زیر را تعیین می‌کنیم:

$$E[z_{im}] = \frac{\alpha'_m \prod_{j=1}^H \prod_{k=1}^{K(j)} (\nu'_{jm}(k))^{\delta(y_{ij}, k)}}{\sum_{n=1}^M \alpha'_n \prod_{j=1}^H \prod_{k=1}^{K(j)} (\nu'_{jn}(k))^{\delta(y_{ij}, k)}} \quad (44-2)$$

M-امین مرحله از مقادیر بیشینه تابع Q در رابطه (۴۲-۲) توسط پارامترهای Θ ، با توجه به ارزش مقادیر مورد انتظار از متغیر $E[z_{im}]$ در E-امین مرحله آن در رابطه (۴۴-۲) برابر است با:

$$\Theta^* = \arg \max Q(\Theta; \Theta') = \arg \max_{\{\alpha_m, v_{jm}\}} \sum_{i=1}^N \sum_{m=1}^M (E[z_{im}] \log \alpha_m + E[z_{im}] \log P_m(y_i | \theta_m)) \quad (45-2)$$

دو بخش سمت راست معادله می‌تواند به طور مستقل بهینه شود. ضریب α_m به راحتی با استفاده از ضرایب لاگرانژ و محدودیت $\sum_m \alpha_m = 1$ به صورت زیر محاسبه شود:

$$\frac{\partial Q(\Theta; \Theta')}{\partial \alpha_m} = \frac{\partial}{\partial \alpha_m} \left(\sum_{i=1}^N \sum_{m=1}^M E[z_{im}] \log \alpha_m + \lambda \left(\sum_{m=1}^M \alpha_m - 1 \right) \right) = 0 \quad (46-2)$$

$$\alpha_m = \frac{\sum_{i=1}^N E[z_{im}]}{\sum_{i=1}^N \sum_{m=1}^M E[z_{im}]} \quad (47-2)$$

به همین ترتیب، با فرض مستقل بودن چگالی مؤلفه شرطی از متغیرهای y_{ij} که در رابطه (۳۳-۲) شرح داده شده است به دست آوردن مقادیر بهینه $(v_{jm}(k))$ ، را تسهیل می‌کنیم. باز هم، محدودیت طبیعی $\sum_k v_{jm}(k) = 1$ و ضریب لاگرانژ λ_{jm} مورد استفاده قرار می‌گیرد:

$$\frac{\partial Q(\Theta; \Theta')}{\partial v_{jm}(k)} = \frac{\partial}{\partial v_{jm}(k)} \left(\sum_{i=1}^N \sum_{m=1}^M E[z_{im}] \log P_m(y_i | \theta_m) + \lambda_{jm} \left(\sum_{k=1}^{K(j)} v_{jm}(k) - 1 \right) \right) = 0 \quad (48-2)$$

$$v_{jm}(k) = \frac{\sum_{i=1}^N \delta(y_{ij}, k) E[z_{im}]}{\sum_{i=1}^N \sum_{k=1}^{K(j)} \delta(y_{ij}, k) E[z_{im}]} \quad (49-2)$$

به طور خلاصه الگوریتم EM با حدس مقادیر $\{\alpha'_1, \dots, \alpha'_M, \theta'_1, \dots, \theta'_M\}$ در چگالی مخلوط شروع می‌شود. بعد از آن، گام‌های E و M آن قدر تکرار می‌شود تا همگرایی معیار پوشش داده شود. پس از آن، با توجه به رابطه (۴۳-۲) مرحله برای متغیر $E[z_{im}]$ محاسبه می‌شود. سپس محاسبه بهترین تخمین برای پارامترها مطابق رابطه (۴۹-۳۳، ۴۷-۲) جهت بیشینه سازی احتمال استفاده می‌شود. معیار همگرایی می‌تواند بر اساس افزایش مقدار در تابع احتمال بین دو نتیجه‌ی M مرحله یا تخصیص پایدار نقاط Y ، یا متناظر آن در X . در واقع معیار پایداری مناسب‌ترین گزینه برای خوشبندی است. راه حل توافقی خوشبندی توسط یک بازیبینی ساده از مقادیر مورد انتظار متغیرهای

$E[z_{im}]$ ، با توجه با اینکه که $E[z_{im}]$ نشان‌دهنده احتمال الگوی y که توسط M -امین مؤلفه مخلوط تولید شده است، به دست می‌آید. پس از همگرایی به دست آمده، یک الگوی y که بزرگ‌ترین مقدار E از برچسب پنهان γ است به مؤلفه تخصیص داده می‌شود. به طور مستقیم، در اصل هر یک از M -مرحله برای تعیین سهم الگوی چگالی مؤلفه مشروط در مقایسه با تخمین حداکثر احتمال با ناظر، از ارزش "Soft" اعضاًی خوش استفاده می‌کند. این روش می‌تواند با استفاده از افزارهای ناقص به تولید خوش‌بندی ترکیبی بپردازد. در برخی از افزارها نتایج خوش‌بندی می‌توان شاهد ظاهر شدن زیرمجموعه‌ای از نمونه‌گیری‌ها یا باز نمونه‌گیری از داده باشیم. برای مثال، یک افزار از یک نمونه خود را انداز ^{۹۹} فقط برچسب‌هایی برای نقاط انتخاب شده را فراهم می‌کند. از این روی، ترکیبی با چنین افزارهایی توسط مجموعه بردارهایی از برچسب خوش نشان داده می‌شود که پتانسیل مفقود شدن مؤلفه‌های را دارد. علاوه بر این، احتمالاً بردارهای متفاوت از برچسب خوش‌بندی مختلف را از دست می‌دهد. وقتی برخی از الگوریتم‌های خوش‌بندی‌ها *Outlier* را به هیچ خوش‌بندی تخصیص نمی‌دهند، اطلاعات ناقص به وجود می‌آیند. خوش‌بندی‌های مختلف در ترکیب‌های گوناگون می‌توانند نقاط مشابه x را به عنوان یک *Outlier* در نظر بگیرند در غیر این صورت، منجر به از دست رفتن مؤلفه‌های بردار y می‌شود. هنوز سناریویی بر جسته دیگری جهت از دست رفتن اطلاعات می‌تواند در ترکیب خوش‌بندی اطلاعات توزیع شده یا ترکیب خوش‌بندی کپی‌های غیر یکسان از یک داده رخ دهد.

پذیرش الگوریتم EM در مورد داده‌های گم شده، یعنی برچسب‌های مفقود شده خوش برای نقاط داده مشابه، امکان‌پذیر است [27]. در این شرایط، هر بردار y از Y به دو بخش مشاهده شده و مفقود شده (y_i^{obs}, y_i^{mis}) تقسیم می‌شود. ادغام داده‌های گم شده منجر به اصلاح جزئی محاسبه مراحل E و M می‌شود. ابتدا، مقادیر مورد انتظار $E[z_{im} | y_i^{obs}, \Theta']$ حالا از مؤلفه‌های مشاهده شده از بردار y استنباط می‌شود، به عنوان مثال رویه رابطه (۴۴-۲) از برچسب‌های شناخته شده گرفته می‌شود:

$$\prod_{j=1}^H \rightarrow \prod_{j:y_i^{obs}} \quad (50-2)$$

^{۹۹} Bootstrap

علاوه بر این، نیاز به محاسبه مقادیر مورد انتظار $E[z_{im}y_i^{mis} | y_i^{obs}, \Theta']$ و جانشین کردن آنها، و همچنین $E[z_{im} | y_i^{obs}, \Theta']$ در M -مرحله برای تخمین مجدد پارامترهای $v_{jm}(k)$ است. جزئیات بیشتر در مورد داده‌های مفقودشده را می‌توان در [45, 25] یافت. شکل زیر شبه کد الگوریتم $K-means$ خوشبندی ترکیبی توسط مدل مخلوط را نشان می‌دهد. در این شبه کد، از الگوریتم $K-means$ برای تولید نتایج اولیه استفاده شده است که می‌توان جای آن از هر الگوریتم دیگری استفاده کرد.

```

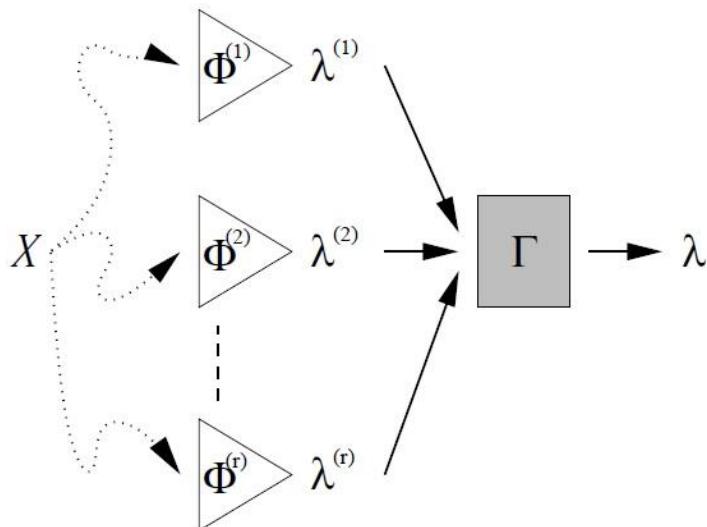
Begin
  for  $i=1$  to  $H$  //  $H$  - number of clusterings
    cluster a dataset  $\pi \leftarrow k\text{-means}(\mathbf{X})$ 
    add partition to the ensemble  $\Pi = \{\Pi, \pi\}$ 
  end
  initialize model parameters  $\Theta = \{\alpha_1, \dots, \alpha_M, \theta_1, \dots, \theta_M\}$ 
  do until convergence criterion is satisfied
    compute expected values  $E[z_{im}], i=1\dots N, m=1\dots M$ 
    compute  $E[z_{im}y_i^{mis}]$  for missing data (if any)
    re-estimate parameters  $v_{jm}(k), j=1\dots H, m=1\dots M, \forall k$ 
  end
   $\pi_c(x_i) =$  index of component of  $Z_i$  with largest expected value,  $i=1\dots N$ 
  return  $\pi_c$  // consensus partition
end

```

شکل ۱۷-۲. زیر شبه کد الگوریتم خوشبندی ترکیبی توسط مدل مخلوط

۲-۲-۳-۲. روش مبتنی بر ابر گراف

این روش در [54] معرفی شده است. در روش مبتنی بر ابر گراف^{۱۰۰}، خوشه‌ها با ابر لبه‌های^{۱۰۱} یک گراف نمایش داده می‌شوند. رأس‌های^{۱۰۲} گراف معادل نمونه‌هایی هستند که باید خوشه‌بندی شوند. مسئله تکه‌تکه کردن این گراف و ایجاد k قسمت متفاوت است که هر قطعه مربوط به یک خوشه می‌شود. سه نوع الگوریتم متفاوت در این خانواده وجود دارد که عبارت‌اند از^{۱۰۳} *CSPA*^{۱۰۴}،^{۱۰۵} *MCLA* و^{۱۰۶} *HGPA*.



شکل ۲. خوشه‌بندی ترکیبی. تابع توافقی Γ برای ترکیب نتایج خوشه‌بندی (λ) استفاده می‌شود [54].

جهت توضیح روش‌های مبتنی بر ابر گراف ابتدا به یک سری از تعاریف پایه می‌پردازیم. مجموعه‌ای از اشیاء، نمونه‌ها یا نقاط است. یک افزار از k شی در خوشه را می‌توان در مجموعه اشیای $\{x_1, x_2, \dots, x_n\}$ یا بردار برچسبⁿ $\lambda \in N^k$ (که N مجموعه اعداد طبیعی است) نشان داد. یک الگوریتم خوشه‌بندی Φ تابعی جهت ارائه بردار برچسب به مجموعه‌ای از اشیاء است. شکل (۱۸-۲) نشان‌دهنده رویه بنیادی اجرای یک خوشه‌بندی ترکیبی است: یک مجموعه^r تایی از برچسب‌های $\lambda^{(1), \dots, r}$ در برچسب λ (برچسب توافقی) با استفاده از تابع توافقی Γ ترکیب

¹⁰⁰ Hyper Graph Partitioning

¹⁰¹ Hyper Edges

¹⁰² Vertices

¹⁰³ Cluster-based Similarity Partitioning Algorithm

¹⁰⁴ HyperGraph-Partitioning Algorithm

¹⁰⁵ Meta-CLustering Algorithm

می‌شوند. بردار ماتریس انتقال با یک بالا نویس (r, \dots, r) در داخل پرانتز نشان داده شده است که برای این بالا نویس برای بیانگر شماره شاخص است و این شماره توان آن بردار / ماتریس انتقال را نشان نمی‌دهد [54].

مرحله اول جهت اجرای روش‌های مبتنی بر ابر گراف این است که مجموعه خوشبندی‌ها را به ابر گرافی مناسب تبدیل کرد. یک ابر گراف شامل رئوس و ابر لبه‌ها می‌شود. لبه در گراف عادی دقیقاً دو رأس را به هم وصل می‌کند. یک ابر لبه به طور کلی همانند یک لبه است که می‌تواند مجموعه‌ای از رئوس را به هم متصل کند. برای هر برچسب بردار $N^{(q)} = \lambda^{(q)}$, یک عضو دودویی در ماتریس $H^{(q)}$ ایجاد می‌شود، با یک ستون برای هر خوشه (که حالا با ابر لبه نمایش داده می‌شود)، شکل (۱۹-۲) مثالی از این روش می‌باشد [54].

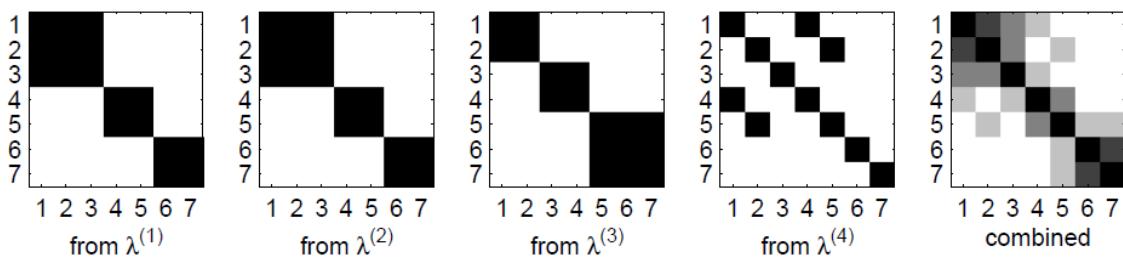
	$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$		$H^{(1)}$	$H^{(2)}$	$H^{(3)}$	$H^{(4)}$					
	h_1	h_2	h_3		v_i	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}	
x_1	1	2	1	1		1	0	0	0	1	0	0	1	0
x_2	1	2	1	2		1	0	0	0	1	0	0	0	1
x_3	1	2	2	?	\Leftrightarrow	1	0	0	0	1	0	0	0	0
x_4	2	3	2	1		0	1	0	0	0	1	0	1	0
x_5	2	3	3	2		0	1	0	0	0	1	0	0	1
x_6	3	1	3	?		0	0	1	1	0	0	0	1	0
x_7	3	1	3	?		0	0	1	1	0	0	0	1	0
					v_7									

شکل ۱۹-۲. نمونه ماتریس $H^{(q)}$ ، جهت تبدیل خوشبندی به ابر گراف [54].

تمام موجودیت‌های یک سطر در ماتریس شاخص اعضای دودویی $H^{(q)}$ ، اگر مربوط به برچسب شناخته شده آن ستون باشد برابر یک و در غیر این صورت برابر با صفر خواهد شد. بلوک چند تیکه‌ی ماتریس $H = H^{(1, \dots, r)} = (H^{(1)} \dots H^{(r)})$ ، ماتریس مجاورت از یک ابر گراف با $\#$ رأس و $\sum_{q=1}^r k^{(q)}$ ابر لبه را تعریف می‌کنند. هر ستون بردار h_a یک ابر لبه h_a را تعریف می‌کند، که یک نشان می‌دهد که، رأس متناظر آن سطر بخشی از ابر لبه است و صفر بودن نشان‌دهنده این است که آن بخشی از ابر لبه نیست. بنابراین ما برای هر یک از خوشه‌ها یک نگاشت به یک ابر لبه و برای مجموعه خوشبندی‌ها یک نگاشت به ابر گراف ارائه کردیم.

CSPA ۱-۲-۳-۲ روش

با یک دید کلی، هر دو شی که در یک خوشه باشند دارای شباهت یک خواهند بود و در غیر این صورت مقدار شباهت آنها صفر است. یک ماتریس شباهت $n \times n$ برای هر خوشه‌بندی می‌تواند بر این اساس ایجاد شود. میانگین ورودی-هوشمندانه^{۱۰۶} از r ماتریس تصویر بهتری از بازده کلی دسته‌بندی r مجموعه در ماتریس شباهت S را نشان می‌دهد. موجودیت‌های S ، کسری از خوشه‌بندی را نشان می‌دهد که در آن دو شی عضو یک خوشه مشابه هستند. ماتریس S را می‌توان به صورت یک ضرب ماتریس اسپارس $S = \frac{1}{r} HH^T$ نشان داد. شکل (۲۰-۲) حالت عمومی ماتریس شباهت بر اساس خوشه‌بندی را برای مثال شکل (۱۹-۲) نشان می‌دهد.



شکل ۲۰-۲. ماتریس شباهت بر اساس خوشه برای مثال شکل (۵-۳). [54]

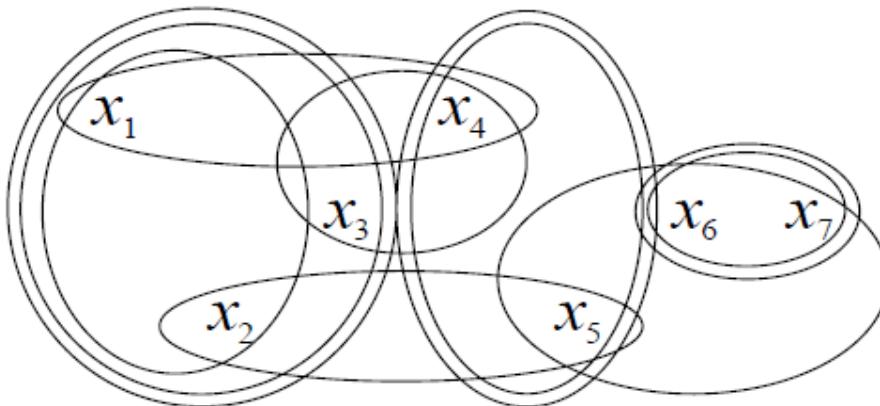
با این روش می‌توان از ماتریس شباهت برای ایجاد مجدد خوشه از اشیای استفاده کرد. در این روش برای تولید گراف (رأس = اشیا، وزن لبه = شباهت) از روش METIS که در [38] ارائه شده است به خاطر خواص خیلی قوی و مقیاس‌پذیر آن، استفاده شده است. روش CSPA یکی از ساده‌ترین روش‌های مکاشفه‌ای^{۱۰۷} جهت ادغام نتایج خوشه‌بندی است ولی پیچیدگی محاسبه و ذخیره‌سازی آن هر دو برابر با درجه دوم $\|$ است که این امر در سایر روش‌های ابر گراف‌ها نزدیک با مقدار خطی n است.

¹⁰⁶ Entry-wise

¹⁰⁷ Heuristic

HGPA .۲-۳-۲-۲-۲-۲ روش

در این روش با فرموله کردن افزایشی ابر گراف توسط قطع حداقل ابر لبه‌ها اقدام به خوشه‌بندی ترکیبی می‌کنیم. این روش الگوریتم افزایشی ابر گراف (HGPA) نامیده می‌شود. در این روش تمام ابر لبه‌ها و رئوس دارای وزن یکسان می‌باشد. باید توجه داشته باشید که این راه حل شامل روابط n طرفه خواهد شد در صورتی که روش CSPA تنها شامل روابط دو به دو می‌باشد.



شکل ۲۱-۲. الگوریتم افزایشی ابر گراف [54].

حال، همانند شکل (۲۱-۲) ما به دنبال جداسازی ابر لبه‌ها برای افزایشی ابر گراف به مؤلفه‌های غیر متصل و تقریباً هم سایز هستیم. باید توجه داشت که اخذ اندازه قابل مقایسه افزایشها در افزایشی گراف‌هایی که بر اساس خوشه‌بندی به دست آمده‌اند یک رویکرد استاندارد جهت اجتناب از افزایشی‌های بی‌اهمیت است [41]. از طرف دیگر معنای این تعریف، این است که اگر خوشه‌های داده طبیعی بسیار نامتعادل باشد، یک رویکرد افزایشی بر اساس گراف مناسب نخواهد بود. در [54] حداکثر عدم تعادل را با حفظ محدودیت $\max_{\ell \in \{1, \dots, k\}} \frac{n_\ell}{n} \leq 1.05$ فرض کردند. افزایشی ابر گراف‌ها در سال‌های اخیر یکی از بهترین حوزه‌های تحقیقاتی بوده است که می‌توان جزئیات برحی از این الگوریتم‌ها را در [38, 65] پیدا کرد. در [54] برای افزایشی روش HMETIS را پیشنهاد شده است [41] دلیل این کار کیفیت بالا افزایشی و مقیاس‌پذیری روش HMETIS می‌باشد. با این حال، باید یادآور شد که افزایشی ابر گراف‌ها به طور کلی دارای هیچ شرایط و قانون خاصی جهت حذف بخشی از ابر لبه‌ها نیست. این بدان معنی است که هیچ حساسیتی جهت وجود تعداد ابر لبه‌ها در یک گروه مشابه بعد از برش وجود ندارد. این برای کاربردهای ما می‌تواند مشکل‌ساز باشد این مسئله را در داده شکل (۱۹-۲) می‌توان شرح داد. برای سادگی کار، اجازه دهید

تا فقط سه ابر لبه برای λ فرض کنیم. دو افزایشندی $\{x_1, x_2, x_7\}, \{x_3, x_4\}, \{x_5, x_6\}$ و $\{x_1, x_7\}, \{x_3, x_4\}, \{x_2, x_5, x_6\}$ هر دو با برش سه ابر لبه ایجاد می‌شود. افزایشندی اول به طور مستقیم بهتر است، به خاطر اینکه $2/3$ از ابر لبه $\{x_1, x_2, x_3\}$ باقی خواهد ماند ولی در روش دوم این مقدار به $1/3$ کاهش پیدا می‌کند. از این روی، در افزایشندی مبتنی بر ابر گراف استاندارد برای تعادل در کیفیت را در حذف هر دو ابر لبه مشابه در نظر می‌گیریم.

۲-۳-۲. روش MCLA

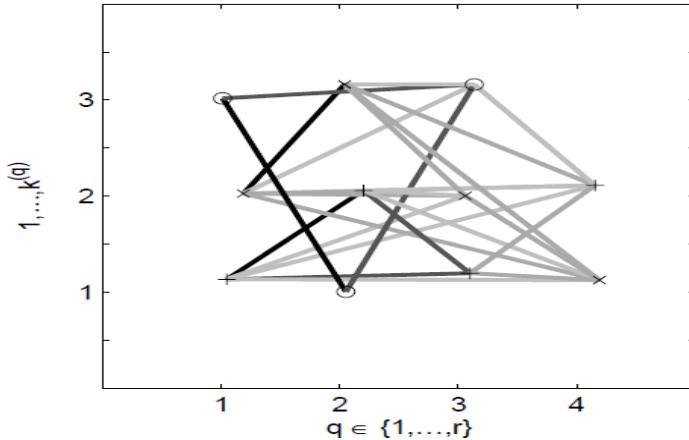
الگوریتم فرا خوشبندی (MCLA) یکی از بهترین روش‌ها در خوشه ترکیبی مبتنی بر ابر گراف است [54]. ایده اصلی الگوریتم فرا خوشبندی بر اساس گروه‌بندی و جداسازی روابط ابر لبه‌ها و تخصیص هر شی به ابر لبه جدا شده است که در آن این مشارکت قویاً دیده می‌شود. ابر لبه‌های مرتبط در نظر گرفته شده برای جداسازی توسط خوشبندی مبتنی بر گراف از ابر لبه‌ها معین می‌شوند. هر خوشه از ابر لبه‌ها به یک ابر خوشه $C^{(M)}$ اشاره می‌کند. جداسازی تعداد ابر لبه‌ها را از $\sum_{q=1}^r k^{(q)}$ به k کاهش می‌دهد. مراحل اجرای الگوریتم فرا خوشبندی به شرح زیر است:

ساخت ابر گراف به عنوان یک گراف بدون جهت دیگر تمام $\sum_{q=1}^r k^{(q)}$ را با h نمایش می‌دهیم (ابر گراف‌های H ، که آن را فرا گراف می‌نامیم. وزن لبه‌ها را متناسب به شباهت بین رئوس در نظر می‌گیریم. در اینجا معیار جاکارت¹⁰⁸ یکی از مناسب‌ترین معیارها برای اندازه‌گیری شباهت هست، از آنجا که آن نسبت بین اشتراک و اجتماع مجموعه‌ای از اشیاء مربوط به دو ابر لبه را نشان می‌دهد. به عبارت دیگر، وزن لبه $w_{a,b}$ بین دو رأس h_a و h_b با معیار جاکارت دودویی مطابق رابطه (۵۱-۲) تعریف می‌شود.

$$w_{a,b} = \frac{h_a h_b}{\|h_a\|_2^2 + \|h_b\|_2^2 - h_a h_b} \quad (51-2)$$

تا زمانی که خوشه‌ها هم پوشانی (خیلی زیاد) نداشته باشند، هیچ لبه‌ای میان رئوس خوشبندی مشابه $H^{(q)}$ وجود نخواهد داشت و بنابراین، فرا گراف r بخشی خواهد بود. شکل (۲۲-۲) الگوریتم فرا خوشبندی مثال شکل (۱۹-۲) است.

¹⁰⁸ Jaccard



شکل ۲-۲. الگوریتم فراخوشبندی

خوشه ابر لبه‌ها^{۱۰۹} در این مرحله ما به دنبال پیدا کردن برچسب‌های سازگار در افزایش‌بندی فرا گراف به k فراخوشه متعادل هستیم. برای این کار [54] روش *METIS* را پیشنهاد کرده است. این نتایج در یک خوشه‌بندی از برداهای h است. هر فراخوشه تقریباً r رأس دارد. از آنجایی که هر رأس در فراخوشه نشان‌دهنده یک برچسب خوشه متمایز است، یک فراخوشه نشان‌دهنده یک گروه از برچسب‌های متناظر است.

جداسازی فراخوشه^{۱۱۰} برای هر یک از k فراخوشه، ابر لبه‌ها برای تبدیل به یک فرا لبه جداسازی می‌شود. هر فرا لبه دارای یک بردار تجمع است که شامل یک ورودی برای هر شی است که سطح تجمع ارتباط فراخوشه را شرح می‌دهد. این سطح برابر با میانگین تمام شاخص‌های بردار h از یک فراخوشه خاص است. هر ورودی صفر و یک به ترتیب نشان‌دهنده قوی‌ترین و ضعیف‌ترین تجمع است.

تخصیص اشیاء^{۱۱۱} در این مرحله، هر شی به فراخوشه‌ای که بیشتر با آن در ارتباط است تخصیص داده می‌شود: به طور خاص، یک شی به فراخوشه‌ای که بالاترین ورودی را در بردار اجماع دارد تخصیص داده می‌شود. روابط به صورت تصادفی شکسته می‌شوند. اطمینان از یک تخصیص، در سهم برنده اجماع منعکس می‌شود (نسبت سهم برنده اجماع به جمع همه اجماع‌های دیگر). باید

¹⁰⁹ Cluster Hyperedges

¹¹⁰ Collapse Meta-clusters

¹¹¹ Compete for Objects

توجه داشت که برای هر فرا خوش نمی توان تضمین داد که حداقل برنده یک شی شود. بنابراین، بیشتر از k برچسب در ترکیب نهایی خوشبندی λ وجود دارد.

شکل (۲۲-۲) نشان دهنده فرا خوش مثال شکل (۱۹-۲) است که در آن $k=3$ ، $r=4$ و $n=2$ می باشد. شکل (۲۲-۲) نشان دهنده یک فرا خوش با چهار قسمت اصلی است. سه فرا خوش توسط سه میل های \circ ، \times و $+$ نشان داده شده است. نشان \circ را به عنوان فرا خوش اول $C_1^{(M)} = \{h_3, h_4, h_9\}$ با $h_1^{(M)} = \{v_5, v_6, v_7\}$ در نظر بگیرید. با جداسازی ابر لبه ها، شی وزن دار فرا لبه $C_1^{(M)}$ در رقابت برای تخصیص بردار اجمع $(0,0,0,0,1/3,1,1)$ حاصل می شود. متعاقباً، فرا خوش $C_1^{(M)}$ در نتایج خوشبندی جامع نشان رئوس / اشیای v_6 و v_7 برنده می شود و بنابراین خوش $C_1 = \{x_6, x_7\}$ در نتایج خوشبندی جامع نشان داده می شود. الگوریتم فرا خوشبندی برای این مثال روی خروجی های $(2,2,2,3,3,1,1)$ که یکی از شش خوشبندی بهینه می باشد و برابر با خوشبندی های $\lambda^{(1)}$ و $\lambda^{(2)}$ است استوار است. عدم قطعیت در برخی از اشیاء به ترتیب در اطمینان $4/3$ ، 1 ، $2/3$ ، 1 ، $1/2$ ، 1 و 1 برای اشیای 1 تا 7 منعکس شده است.

۳-۲-۳-۲. روش های مبتنی بر ماتریس همبستگی

Input: D – the input data set N points
 B – number of partitions to be combined
 M – number of clusters in the final partition, σ
 k – number of clusters in the components of the combination
 Γ – a similarity-based clustering algorithm
 for $j=1$ to B
 Draw a random pseudosample X_j
 Cluster the sample X_j : $\pi(i) \leftarrow K\text{-means}(\{X_j\})$
 Update similarity values (co-association matrix) for all patterns in X_j
 end
 Combine partitions via chosen Γ : $\sigma \leftarrow \Gamma(P)$
 Validate final partition, σ (optional)
 return $\sigma //$ consensus partition

شکل ۲۳-۲. الگوریتم خوشبندی ترکیبی مبتنی بر ماتریس همبستگی و با استفاده از توابع توافقی مختلف مبتنی بر شباهت

در روش ماتریس همبستگی¹¹² شباهت بین نقاط (مقادیر همبستگی)، می‌تواند با تعداد خوش‌های به اشتراک گذاشته شده بین دو نقطه، در همه افزارهای یک ترکیب، تخمین زده شود. ساختار این نوع از الگوریتم‌های خوش‌بندی ترکیبی در شکل ۲-۳ نشان داده شده است.

۱-۲-۳-۲-۳-۲. الگوریتم‌های سلسله مراتبی تراکمی

فرض کنید مجموعه داده D شامل N نقطه (نمونه) در فضای d بعدی است. داده‌های ورودی را می‌توان به صورت یک ماتریس الگوی $N \times d$ و یا یک ماتریس عدم تشابه $N \times N$ در نظر گرفت. فرض کنید $\{X_1 \dots X_B\}$ مجموعه‌ی زیرمجموعه‌ی نمونه‌های نمونه‌های ماست که از نمونه‌های اولیه استخراج شده‌اند. هر یک از الگوریتم‌های انتخابی هنگامی که بر روی زیرمجموعه‌ی نمونه‌های موجود در X اجرا شوند نتایج $\{\pi_1 \dots \pi_B\} = \Pi$ را تولید می‌کنند. هر π_i مجموعه‌ای از خوش‌های است. یا به عبارت دیگر $\{\pi_i = \{C_1^i, C_2^i, \dots, C_{k(i)}^i\}\}$ و به ازای هر π_i داریم $X_i = C_1^i \cup C_2^i \dots \cup C_{k(i)}^i$ به طوری که (i) تعداد خوش‌ها در i امین خوش‌بندی است. اولین یک الگوریتم پایه (برای مثال $K-means$) را بر روی $\{X_1 \dots X_B\}$ اجرا می‌کنیم تا بتوانیم با استفاده از π_i ‌های تولید شده ماتریس همبستگی را به صورت زیر به دست آوریم:

$$Sim(x, y) = \frac{1}{b} \sum_{i=1}^B \delta(\pi_i(x), \pi_i(y)) \quad (52-2)$$

$$\delta(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases} \quad (53-2)$$

در رابطه ۵۲-۲، تابع $\delta(\pi_i(x), \pi_i(y))$ در صورتی که دو عنصر x و y در خوش‌بندی i در یک خوش‌بندی باشند، مقدار یک و در غیر این صورت مقدار صفر برمی‌گرداند. مقدار پارامتر B نمایانگر تعداد زیرمجموعه‌های است و یا به بیان دیگر تعداد دفعات تکرار الگوریتم پایه است. معمولاً از الگوریتم‌های سلسله مراتبی پیوندی (منفرد، کامل، میانگین و بخشی) برای ترکیب از روی ماتریس همبستگی استفاده می‌شود [33]. سه اشکال اصلی روش‌های مبتنی بر ماتریس همبستگی عبارت‌اند از:

۱- یک پیچیدگی محاسباتی درجه دوم در تعداد الگوها و ویژگی‌ها ($O(kN^2d^2)$ دارند.

¹¹² Co-association Matrix

۲- هیچ راهنمایی برای اینکه کدام الگوریتم خوشبندی باید به کاربرده شود، وجود ندارد. به عنوان مثال پیوندی منفرد یا پیوندی کامل.

یک ترکیب با یک تعداد کوچک از افزارها، ممکن است یک تخمین مطمئن از مقادیر همبستگی را فراهم نکند.

۲-۳-۲-۳-۲. الگوریتم افزایشی گراف با تکرار

در روش‌های معمول خوشبندی ترکیبی پس از تشکیل ماتریس همبستگی حتماً باید تعداد خوشبندی نتیجه ترکیب را تعیین کرد. روشی که توسط [50] به تازگی معرفی شده است، راه حلی جهت تشخیص بهترین تعداد خوشبندی نهایی را به صورت خودکار معرفی می‌کند. در این روش ابتدا ماتریس همبستگی که در اینجا به آن ماتریس قضاوت می‌گویند را تشکیل می‌دهیم، سپس توسط الگوریتم افزایشی تکراری مبتنی بر گراف به صورت خودکار تعداد خوشبندی‌های نهایی را تشکیل می‌دهیم. نتایج تجربی در [50] نشان می‌دهد که همیشه فراهم کردن نتایج بهینه با تعیین تعداد خوشبندی نهایی به صورت ثابت امکان‌پذیر نیست. در این روش یک فرآیند افزایشی گراف تکراری هر بار مقادیر ماتریس همبستگی J را یک درجه کاهش می‌دهد به طوری که به تدریج اتصال میان نقاط داده شکسته شود. افزایشی گراف اصلی G (که گراف معادل ماتریس همبستگی J است) را به زیر گراف‌ها تقسیم می‌کند. برای تشخیص تعداد خوشبندی نهایی کافی است تا تعداد زیر گراف‌ها را بشماریم. الگوریتم افزایشی گراف با تکرار دو مرحله کلی دارد، ابتدا کاهش درجه ماتریس و سپس شمارش تعداد زیر گراف‌ها.

کاهش ماتریس‌ها یک رویه جهت کاهش ماتریس همبستگی J به یک درجه کمتر در یک مرحله تکراری است، این رویه تا جایی تکرار می‌شود تا تمام موجودیت‌ها شروع به صفر شدن کنند. این رویه به صورت رابطه زیر نمایش داده می‌شود.

$$t_{ij}^{new} = \begin{cases} t_{ij}^{previous} - 1 & \text{if } t_{ij}^{previous} > 0, \\ 0 & \text{otherwise} \end{cases}, t_{ij}^{new} \in t_{ij}^{new}, t_{ij}^{previous} \in t_{ij}^{previous} \quad (54-2)$$

در این رابطه ماتریس J^{new} کسر یک بر روی هر ورودی در ماتریس پیشین $J^{previous}$ است. رویه کاهش ماتریس پیشنهادشده به تدریج ارتباطات سمت بین نقاط داده را می‌شکند.

شمارش زیر گراف‌ها یک رویه برای شمارش تعداد زیر گراف‌های متصل در هر ماتریس کاهش یافته است و برای ارزیابی شرایط خوشبندی در طول فرآیند کاهش استفاده می‌شود. الگوریتم پیمایشی گراف برای شمارش گراف G^{new} از ماتریس J^{new} پیاده‌سازی شده است، به عنوان مثال الگوریتم BSF^{113} که به صورت زیر محاسبه می‌شود:

$$G_{subgraphs} = BSF_traversal(J^{new}) \quad (55-2)$$

ماتریس همبستگی J که از $k-1$ متریک مشاهده‌ای جمع‌آوری شده است به عنوان ورودی فرآیند افزایش‌بندی گراف عمل می‌کند. در طول هر مرحله از کاهش $J^{previous}$ ، یک ماتریس جدید J^{new} و گراف مجاورت G^{new} آن تولید می‌شود. هر G^{new} شامل تعدادی از زیر گراف‌ها می‌باشد. تعداد این زیر گراف‌ها در واقع تعداد خوشه در نتیجه نهایی می‌باشد. فرآیند افزایش‌بندی تا وقتی که تمامی ورودی‌های $J^{previous}$ صفر شوند ادامه پیدا می‌کند. الگوریتم افزایش‌بندی گراف با تکرار در شبه کد شکل (۲۴-۲) نشان داده شده است.

Input: A judgment matrix J

$n = 0;$

$G^{new}[0] = J$. BSF_traversal();

$J^{previous} = J$;

Do

$n = n + 1$;

$J^{new} = \text{Decreasing}(J^{previous})$

$G_{subgraphs}[n] = \text{BSF_traversal}(J^{new})$

$\text{ClusterNumber}[n] = G_{subgraphs}[n]. \text{getSubGraphNumber}();$

$J^{previous} = J^{new}$;

Until (all entries in $J^{previous}$ previous are 0)

Return $\text{ClusterNumber}[n]$ and $G_{subgraphs}[n]$;

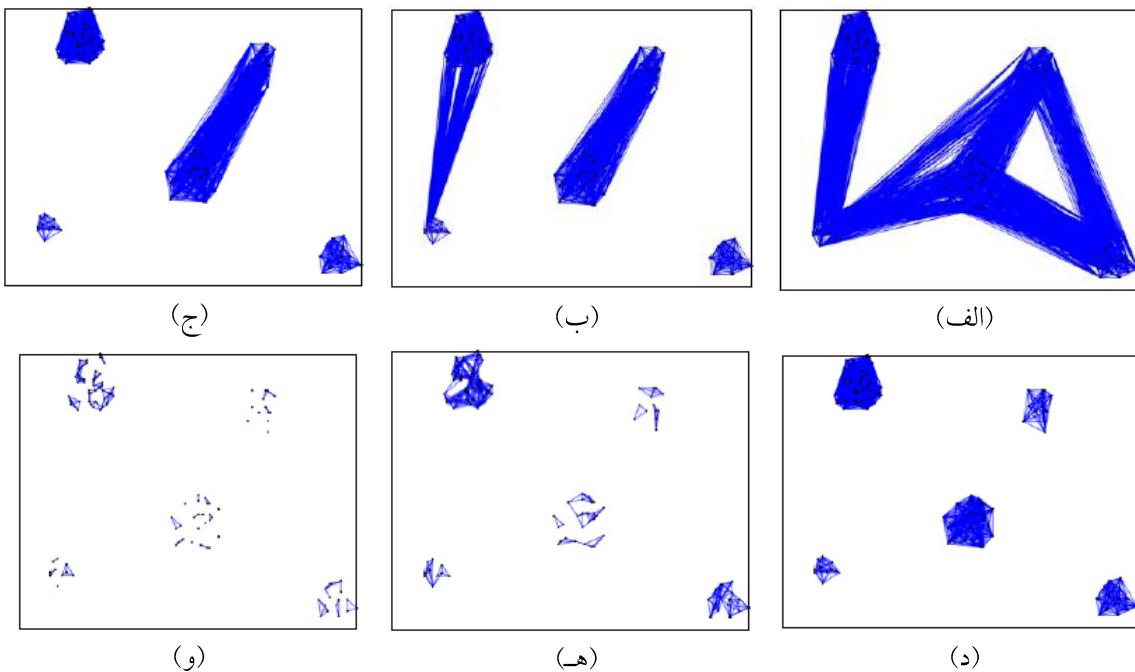
Output: An array of cluster numbers- $\text{ClusterNumber}[n]$ and a set of $sub-graphs - G_{subgraphs}$

شکل (۲۴-۲). الگوریتم افزایش‌بندی با تکرار

شکل (۲۵-۲) قسمتی از فرآیند کاهش و محاسبه را به تصویر می‌کشد. گراف مجاورت ماتریس همبستگی اصلی قسمت (الف) شکل (۲۵-۲) نشان داده شده است. در این مورد نشان داده شده

¹¹³ Breadth first search

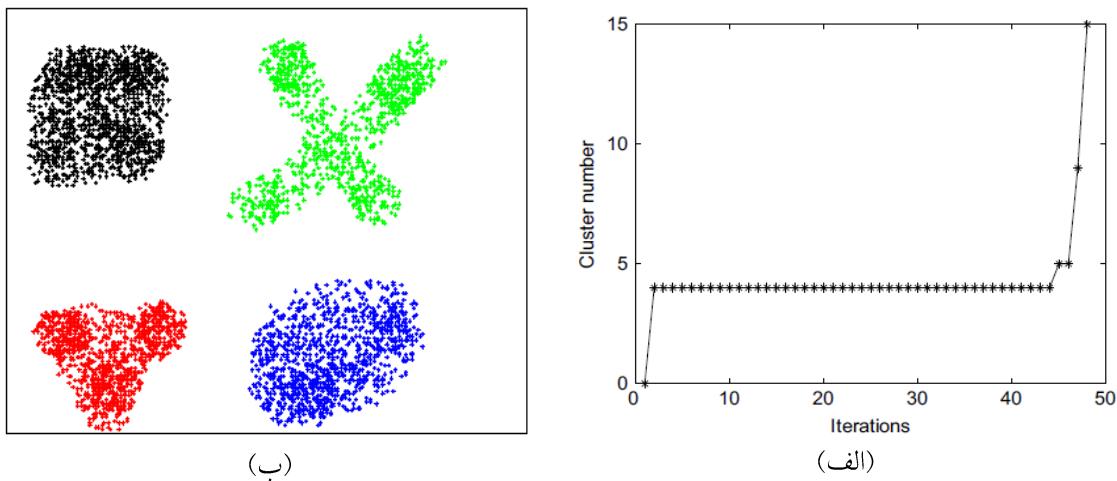
است که تمامی نقاط داده متصل شده‌اند و بنابراین آن‌ها متعلق به تنها یک خوش‌هه هستند. در بخش (ب) شکل (۲۵-۲) سه زیر گراف پس از چند تکرار افرا بندی گراف نشان داده شده است، که در آن اتصالات بین سه زیر گراف قطع شده است. این بدان مفهوم می‌باشد که اتصالات این نقاط داده به طور مقایسه‌ای ضعیف‌تر از سایر نقاط متصل می‌باشد. از این‌رو در این مرحله سه خوش‌هه شناسایی شده است. در قسمت (ج) و (د) شکل (۲۵-۲) چهار و پنج خوش‌هه بعد از تکرا الگوریتم افزایش‌بندی گراف پیدا شده‌اند. اگر این فرآیند مطابق بخش (ه) و (و) ادامه پیدا کند می‌توان تعداد بیشتری از زیر گراف‌ها را پیدا کرد. به عنوان نتیجه هر مرحله کاهش و شمارش اتصالات زیر گراف‌ها را قطع کرده و یک تعداد از نتایج خوش‌بندی پیدا می‌شود. با این وجود، باید به این سؤال پاسخ داد که چگونه نتایج نهایی خوش‌بندی و تعداد دلخواه خوش‌بندی به دست می‌آید.



شکل ۲۵-۲. نمایش گراف مجاورت در مراحل کاهش درجه ماتریس و شمارش آن [۵۰].

در این روش، هر تکرار از فرآیند افزایش‌بندی گراف تعدادی از زیر گراف‌ها را تولید می‌کند، و تعداد زیر گراف‌ها دقیقاً برابر با تعداد خوش‌ههای آن تکرار می‌باشد. در توزیع عددی زیر گراف همان‌طور که در شکل (۲۶-۲) نشان داده شده است تعداد زیر گراف‌ها ممکن است برای تعدادی از تکرارها ثابت باقی بماند سپس یک تغییر قابل‌لمس را پس از مرحله ثبات شاهد هستیم. تعداد مطلوب خوش‌هه بر این اساس به صورت باثبات‌ترین تعداد زیر گراف در توزیع تعریف شده است، به این دلیل که تحت این تعداد زیر گراف اتصالات نقاط داده در زیر گراف‌ها سخت‌ترین و قوی‌ترین در مقابله با شکستن هستند. شکل (۲۶-۲) مثالی از توزیع تعداد زیر گراف و نتایج ترکیب نهایی را نشان می‌دهد.

در شکل (۲۶-۲) بدیهی است که عدد چهار تعداد مطلوب می‌باشد که هم در شکل پایداری توزیع (الف) و هم در تعداد خوشه‌ها (ب) نمایان است.



شکل ۲۶-۲. مثال روند تغییر تعداد خوشه [50].

الگوریتم افزاینده گراف با تکرار یک راه حل تطبیق‌پذیر را دنبال می‌کند زیر آن می‌تواند به راحتی با الگوریتم‌های خوشه‌بندی دیگر برای به دست آوردن نتایج قابل‌اتکا و شناسایی تعداد خوشه‌بندی دلخواه کار کند. شبه کد شکل (۲۷-۲) یک جریان کار عمومی برای پیاده‌سازی الگوریتم افزاینده گراف با تکرار در الگوریتم‌های دیگر را به تصویر می‌کشد.

Input: A data set $Y = \{y_1, \dots, y_M\}$ in n-dimensional space

1. Using OCA to cluster the data set Y with cluster number ranging from 2 to k;
2. Every clustering result can be represented by a vector $L = [l_1, l_2, \dots, l_M]$, where

$$l_i = l_j = c \text{ if both } y_i \text{ and } y_j \text{ belongs to the } c-th \text{ cluster;} \\$$
3. An observation matrix O can be computed from the vector L according to

$$o_{ij} = \begin{cases} 1 & l_i = l_j (i \neq j) \\ 0 & otherwise \end{cases};$$
4. A judgment matrix J can be obtained as the sum of $k-1$ observation matrices O,
i.e., $J = \sum_{c=2}^K O_c$;
5. The adjacency graph of J is iteratively partitioned so as to identify the desired cluster number and clustering results according to the distribution of the number of sub-graphs.

Output: The desired cluster number and final clustering result.

شکل ۲۷-۲. جریان کار عمومی برای پیاده‌سازی الگوریتم افزاینده گراف

۳-۳-۲. الگوریتم‌های خوشبندی ترکیبی کامل

الگوریتم‌های خوشبندی ترکیبی کامل^{۱۱۴} روشی کاربردی و ساده جهت ایجاد خوشبندی ترکیبی است. در این روش ما تمامی نتایج اولیه به دست آمده را بدون هیچ شرطی باهم توسط یکتابع توافقی ادغام می‌کنیم. روش انباست مدارک^{۱۱۵} (EAC) یکی از روش‌های خوشبندی کامل است که توسط فرد و چین معرفی شد است. در این روش معمولاً برای ایجاد نتایج اولیه خوشبندی از استفاده می‌شود و تابع توافقی آن روش الگوریتم‌های سلسله مراتبی تراکمی می‌باشد. در روش انباست مدارک ابتدا بر اساس رابطه (۵۶-۲) ماتریس همبستگی را تشکیل می‌دهیم و سپس روی این ماتریس همانند یک ماتریس شباهت/تفاوت ساده توسط الگوریتم سلسله مراتبی یک خوشبندی انجام می‌دهیم [۱۹].

$$C(i, j) = \frac{n_{i,j}}{m_{i,j}} \quad (56-2)$$

در رابطه (۵۶-۲) پارامتر $n_{i,j}$ تعداد دفعاتی است که جفت نمونه‌های i و j باهم در یک خوشبندی شده‌اند و $m_{i,j}$ تعداد نمونه‌برداری‌هایی است که هر دوی این جفت نمونه‌ها به طور همزمان در آن ظاهر شده‌اند.

۴-۲. خوشبندی ترکیبی مبتنی بر انتخاب

اولین بار فرن و لین در [۲۳] خوشبندی ترکیبی مبتنی بر انتخاب^{۱۱۶} را با این عنوان بر اساس ایده‌ی حادجی‌تودوروو^{۱۱۷} و همکاران [۸۴] معرفی کردند. که در آن این نظریه مطرح می‌شود که تعدادی از افزارها و یا خوشبندی‌های ارزیابی شده در مجموعه افزارهای خوشبندی ترکیبی می‌تواند جواب نهایی بهتری نسبت به ترکیب کامل تمامی اعضای مجموعه تولید کند. در این روش علاوه بر چالش‌های پیشین مطرح شده در خوشبندی ترکیبی (نحوی ساخت نتایج اولیه و نحوی ترکیب آنها برای تولید نتیجه‌ی نهایی) دو مسئله‌ی ارزیابی نتایج اولیه و راهکار^{۱۱۸} انتخاب مجموعه منتخب از

¹¹⁴ Full Ensemble

¹¹⁵ Evidence Accumulation Clustering

¹¹⁶ Cluster Ensemble Selection

¹¹⁷ Hadjitarov S. T.

¹¹⁸ Strategy

نتایج ارزیابی شده مطرح خواهد شد. در سال‌های اخیر برخی از مقالات راهکارهایی جهت حل دو چالش مطرح شده بیان کرده‌اند. اولین راه حل در این روش توسط فرن و لین در [23] بیان شده است که این تحقیق آن را خوشبندی ترکیبی مبتنی بر انتخاب فرن و لین نام‌گذاری می‌کند. علاوه بر آن چندین روش دیگر ارزیابی و انتخاب نیز مطرح شده است که ما آن‌ها را به ترتیب زمان ارائه بررسی خواهیم کرد.

۴-۲. خوشبندی ترکیبی مبتنی بر انتخاب فرن و لین

در این روش برای خوشبندی ترکیبی از زیرمجموعه‌ی موثرتری از افزارهای اولیه در ترکیب نهایی استفاده می‌شود. اگر چه تعداد اعضای شرکت‌کننده در ترکیب نهایی در این روش کمتر از یک خوشبندی ترکیبی کامل است، به دلیل انتخاب افزارها با کارایی بالاتر، نتایج نهایی بهبود می‌یابند. پارامترهایی که در این روش مورد توجه قرار گرفته‌اند، عبارت‌اند از: کیفیت و پراکندگی.

در یادگیری با ناظر مفاهیم کیفیت و پراکندگی خوش‌تعریف هستند، کیفیت دقت اعضای ترکیب و پراکندگی تفاوت بین پیش‌بینی‌های انجام‌شده توسط اعضای ترکیب را اندازه‌گیری می‌کنند ولی این مفاهیم در یادگیری بدون ناظر به صورت واضح تعریف‌نشده‌اند [23]. در این بخش رویکردهای روش فرن و لین را جهت حل چالش‌های پیش روی شرح می‌دهیم که در آن ابتدا روش اندازه‌گیری کیفیت و پراکندگی خوشبندی را بیان می‌کنیم و سپس برای هر دو معیار فوق یک راهکار انتخاب توصیف می‌کنیم.

۱-۱-۴-۲. تعریف معیار کیفیت در روش فرن و لین

در روش‌های بدون ناظر، ما هیچ تابع هدف خارجی همانند دقت برای اندازه‌گیری کیفیت راه حل‌های خوشبندی نداریم. در ادبیات خوشبندی، به طور معمول، یک برچسب کلاس به عنوان جانشین برای ساختار زیربنایی درست استفاده می‌شود و سپس بر مبنای این که چگونه این راه حل برچسب‌های کلاس را بازیابی می‌کند کیفیت خوشبندی اندازه‌گیری می‌شود. این روش در خوشبندی ترکیبی مبتنی بر انتخاب نمی‌تواند استفاده شود به خاطر این که اطلاعات با ناظر همانند برچسب کلاس نمی‌تواند در فرآیند خوشبندی استفاده شود. در اینجا، ما یک معیار داخلی کیفیت بر اساس تابع هدف معرفی شده توسط استرل و گاوش [54] برای طراحی تابع توافقی پیشنهاد می‌کنیم.

به طور خاص، یک ترکیب E از r راه حل خوشبندی توسط $\{C_1, C_2, \dots, C_r\}$ نشان داده شده است، به دنبال پیدا کردن خوشبندی توافقی که این معیار را بیشینه می‌کند:

$$SNMI(C, E) = \sum_{i=1}^r NMI(C, C_i) \quad (43-2)$$

در اینجا $NMI(C, C_i)$ معیار اطلاعات متقابل نرمال شده بین خوشبندی‌های C و C_i می‌باشد. اگر دو خوشبندی کاملاً مستقل از یکدیگر باشند مقدار NMI برابر با صفر خواهد شد. در مقابل، اگر دو خوشبندی کاملاً مشابه باشد آنگاه مقدار NMI برابر با یک خواهد شد. در اینجا تابع هدف ما برابر با جمع اطلاعات متقابل نرمال شده ($SNMI$) است. در این روش برای یک مجموعه از خوشبندی‌های $L = \{C_1, C_2, \dots, C_r\}$ جهت اندازه‌گیری کیفیت هر خوشبندی C_i از رابطه $SNMI(C_i, L)$ استفاده می‌کنیم.

۲-۱-۴-۲. تعریف معیار پراکندگی در روش فرن و لین

چندین معیار مختلف برای خوشبندی ترکیبی پیشنهاد شده است. فرن و لین [23] معیاری که توسط فرن و برودلی [22] ارائه شده است را به عنوان معیار پراکندگی پیشنهاد داده‌اند که بر پایه اطلاعات متقابل نرمال شده دوتایی میان راه حل خوشبندی‌ها است. به طور خاص، ما تشابه دو جفت خوشبندی را به صورت $NMI(C_i, C_j)$ اندازه‌گرفته و جمع تمامی تشابهات دوتایی را در ترکیب محاسبه کرده و به عنوان معیار پراکندگی ترکیب در نظر می‌گیریم. هر چند ارزش $\sum_{i \neq j, C_i, C_j \in E} NMI(C_i, C_j)$ کمتر باشد پراکندگی بیشتر است. معیار پراکندگی بالا به این علت انتخاب شده است چون نشان داده که در عملکرد خوشبندی ترکیبی تأثیر می‌گذارد. باید متذکر شد که متد انتخابی [23] خود را محدود به هیچ معیار پراکندگی خاصی نمی‌کند.

۲-۱-۴-۳. راهکار انتخاب خوشبندی برای تشکیل نتیجه نهایی در روش فرن و لین

کیفیت: در اولین مرحله از این روش، از معیارهای تعریف‌شده پیشین برای راهنمایی انتخابمان استفاده می‌کنیم و تنها این راه حل‌ها است که دارای کیفیت بالایی در ترکیب می‌باشند [23]. به طور خاص، اگر مجموعه راه حل‌های خوشبندی L به ما داده شده باشد، این راهکار تمامی راه حل‌های خوشبندی را در L رتبه‌بندی کرده که این عمل را بر پایه کیفیت آنها با استفاده از $SNMI(C, L)$ انجام می‌دهد و K راه حل با رتبه بالا را برای حضور در ترکیب انتخاب می‌کند، که K اندازه دلخواه

ترکیب می‌باشد. ما این راهکار را کیفیت می‌نامیم. توجه کنید که اگر یک راهکار مقدار $SNMI$ بالای داشته باشد، به طور مفهومی، این راه حل سازگاری بالایی را با روند عمومی که توسط کل مجموعه نشان داده شده است، دارا است. از طرف دیگر، راه حل‌های خوشبندی بالارزش $SNMI$ پایین می‌توانند به عنوان *outliers* مجموعه در نظر گرفته شوند و ممکن است که برای حضور در ترکیب مفید باشند. به طور کلی، ترکیب‌های انتخاب‌شده توسط کیفیت می‌توانند افزونگی بالایی را در راه حل‌های انتخاب‌شده داشته باشد.

پراکندگی، در مقابل، این روش به دنبال راهکارهای انتخابی می‌گردد که مقدار پراکندگی را بیشینه کند. اینجا می‌توان سنگین‌ترین گراف k رأسی¹¹⁹ را به عنوان یک مشکل در نظر گرفت. به طور خاص، راه حل خوشبندی در مجموعه به عنوان رئوس در گراف کاملاً متصل، نشان داده می‌شود و مقدار پراکندگی هر جفت آنها ($NMI - 1$) به عنوان وزن لبه‌هایی که به رئوس متصل‌اند، اختصاص داده می‌شوند. انتخاب یک ترکیب به اندازه‌ی K ، با پراکندگی بیشینه می‌تواند به وسیله یافتن یک زیر گراف با K رأس که وزن لبه‌های آن بیشینه شده می‌باشد (که همان سنگین‌ترین گراف K رأسی می‌باشد) به دست آید. با این وجود، این مشکل از درجه سختی NP برخوردار است [39]. در [23] از یک راهکار ساده حریصانه که به صورت زیر توصیف شده است استفاده می‌کنیم.

ما با یک ترکیب E که شامل یک راه حل باکیفیت بسیار بالا می‌باشد شروع می‌کنیم (همان طور که توسط $SNMI$ اندازه‌گیری شده است). سپس آن به صورت افزایشی در هر زمان یک راه حل از کتابخانه برای اضافه کردن به E انتخاب می‌کند. این عمل به این دلیل انجام می‌شود که ترکیب نهایی بیشترین پراکندگی را که همان کمترین مقدار $(\sum_{i \neq j, C_i, C_j \in E} NMI(C_i, C_j))$ می‌باشد، داشته باشد. این فرآیند تکرار می‌شود تا اینکه ما به ترکیب دلخواه‌مان به سایز K برسیم. در ادامه ما این راهکار را با عنوان پراکندگی می‌شناسیم.

در ادبیات موضوعی در [23]، الگوریتم مکاشفه‌ای مختلفی برای تولید راه حل‌های خوشبندی مختلفی برای خوشبندی ترکیبی پیشنهاد شده است و به صورت عمومی این اعتقاد وجود دارد که متنوع کردن خوشبندی ترکیبی تأثیر سودمندی خواهد داشت. زیرا اشتباهاتی که توسط اعضای مختلف ترکیب انجام می‌شود ممکن است که هم‌دیگر را حذف کنند. راهکار پراکندگی که در اینجا صحبت

¹¹⁹ Heaviest K-vertex subgraph

شد از این فلسفه پیروی کرده و به طور صریح به دنبال زیرمجموعه‌های با پراکندگی بالا از مجموعه، برای شکل دادن ترکیب می‌گردد. توجه کنید که مشکل بالقوه در مورد این روش (متد) این است که ممکن است منجر به شمول برخی راه حل‌ها باکیفیت پایین در ترکیب شود.

۲-۴-۲. الگوریتم هوشمند طبقه‌بندی مجموعه داده‌ها

یکی دیگر از روش‌هایی که در خوشبندی ترکیبی مبتنی بر انتخاب ارائه شده است روش عظیمی و همکاران می‌باشد [7]. در این روش از مفهوم پراکندگی برای هوشمند نمودن خوشبندی ترکیبی استفاده شده است و به صورت پویا اقدام به انتخاب زیرمجموعه بهینه‌ای از نتایج اولیه در ترکیب نهایی می‌شود، ابتدا یک خوشبندی ترکیبی ساده انجام می‌شود و سپس این روش میزان شباهت تمام نتایج خوشبندی‌های اولیه را نسبت به جواب به دست آمده ارزیابی می‌کند و سعی در طبقه‌بندی^{۱۲۰} مجموعه داده‌ها به سه مجموعه داده راحت^{۱۲۱}، معمولی^{۱۲۲} و سخت^{۱۲۳} می‌کند. در این طبقه‌بندی، مجموعه داده راحت به مجموعه داده‌ای اطلاق می‌شود که خوشبندی‌های اولیه تفاوت چندانی با خوشبندی ترکیبی به دست آمده نداشته باشند. به این معنی که هر خوشبندی ساده بتواند تقریباً مانند خوشبندی ترکیبی نتایج مشابه ای ارائه کند. مجموعه داده معمولی به مجموعه داده‌ای اطلاق می‌شود که خوشبندی‌های اولیه نه تفاوت چندانی و نه تشابه چندانی با نتایج خوشبندی ترکیبی به دست آمده دارند. مجموعه داده سخت به مجموعه داده‌ای اطلاق می‌شود که خوشبندی‌های اولیه تشابه چندانی با خوشبندی ترکیبی به دست آمده نداشته باشند. این رویداد نشان می‌دهد که داده‌های مجموعه مورد نظر کاملاً دارای مرزهای مشترک هستند و روش‌های ساده و معمولی خوشبندی همانند روش‌های پیچیده و قدرتمند خوشبندی ترکیبی قادر به جداسازی نمونه‌ها نمی‌باشند. سپس کل نتایج خوشبندی‌های اولیه به چهار زیرمجموعه متفاوت بر اساس میزان تطبیق دقت‌شان با نتایج خوشبندی ترکیبی ساده تقسیم می‌شوند و بر اساس رده^{۱۲۴} هر مجموعه داده (راحت، معمولی و سخت) اقدام به انتخاب یکی از این زیرمجموعه‌ها برای ترکیب و به دست آوردن نتیجه نهایی می‌کنیم [5, 7].

¹²⁰ Classification

¹²¹ Easy

¹²² Intermediate

¹²³ Hard

¹²⁴ Class

۳-۴-۲. خوشبندی ترکیبی طیفی مبتنی بر انتخاب بر اساس شباهت

این الگوریتم توسط ژیا و همکاران ارائه شده است که در آن مؤلفه‌های خوشبندی که توسط الگوریتم‌های خوشبندی طیفی تولید می‌شوند، قادر به ایجاد کمیته‌های پراکنده‌گی^{۱۲۵} خواهند بود. همچنین در این روش، پارامتر مقیاس گذاری تصادفی در تقریب نیستروم^{۱۲۶} و مقادیر تصادفی اولیه الگوریتم *K-means*^{۱۲۷} برای آش忿^{۱۲۸} الگوریتم خوشبندی طیفی در تولید مؤلفه‌های یک ترکیب مورد استفاده قرار می‌گیرند. علاوه بر آن یک معیار بر اساس ترکیب معیارهای پراکنده‌گی و کیفیت برای ارزیابی کیفیت تمامی نتایج به دست آمده پیشنهاد شده است و یک روش جدید انتخاب خوش برا اساس قانون نزدیک‌ترین همسایه جهت انتخاب نتایج اولیه معرفی می‌شود [86].

۱-۳-۴-۲. معیار ارزیابی *Sim* در روش پیشنهادی ژیا

در روش ژیا و همکاران مطابق رابطه زیر از تفاصل عدد یک از اطلاعات متقابل نرمال شده به عنوان پراکنده‌گی استفاده شده است [86].

$$Div = 1 - NMI \quad (57-2)$$

در رابطه بالا هر چه معیار *Div* بزرگ‌تر باشد پراکنده‌گی بین دو خوشبندی مورد ارزیابی بیشتر خواهد بود. این معیار برای هر دو خوشبندی مساوی برابر با صفر می‌باشد. از طرف دیگر در ساخت یک ترکیب مناسب باید معیار دقت نیز در نظر گرفته شود تا نتایج نامطلوب، کارایی کل سیستم را مورد تأثیر قرار ندهند. از این روی روش پیشنهادی ژیا یکتابع جدید برای ارزیابی هم زمان دقت و پراکنده‌گی مطابق با رابطه زیر با عنوان *Sim* پیشنهاد کرده است [86].

$$Sim = -(Div \times \ln(Div) + (1 - Div) \times \ln(1 - Div)) \quad (58-2)$$

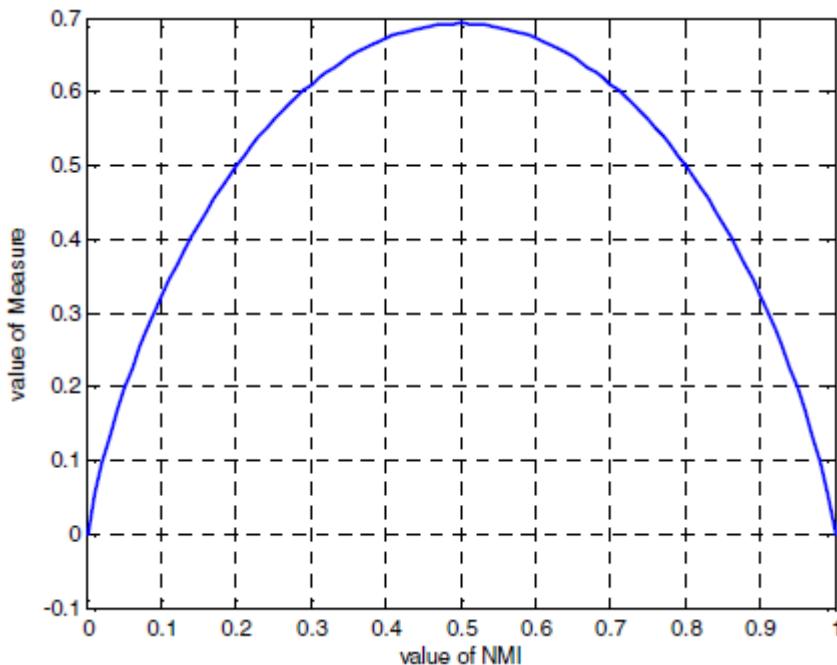
شکل (۲۸-۲) نشان‌دهنده‌ی منحنی تابع *Sim* بر اساس *NMI* در بازه بین صفر الی یک می‌باشد. این شکل نشان می‌دهد که هر گاه مقدار *NMI* یک و یا صفر باشد مقدار *Sim* صفر است. این تحقیق خوشبندی‌هایی با مقدار *NMI* صفر را به عنوان نویز در نظر می‌گیرد و در صورت برابر بودن نتایج

¹²⁵ Diverse Committees

¹²⁶ Nyström approximation

¹²⁷ Perturb

دو خوشبندی (که مقدار NMI آن دو برابر یک می‌باشد) در این روش یکی از آن دو نتیجه مورد استفاده قرار می‌گیرد. مطابق تعاریف بالا تنوع نتایج اولیه انتخاب شده در این روش در حد متوسط^{۱۲۸} می‌باشد [86].



شکل ۲-۲. گراف تابع Sim در بازه بین صفر و یک [86].

۲-۳-۴-۲. انتخاب خوشبندی بر اساس قانون نزدیکترین همسایه در روش ژیا

در خوشبندی مبتنی بر انتخاب پیشنهادی این تحقیق دو فرآیند انجام می‌شود. ابتدا افزایشندی یک مجموعه خوشبندی به تعداد زیرمجموعه‌های همگن تقسیم می‌شود و سپس از هر یک از آنها یک خوشبندی نماینده انتخاب می‌شود. افزایشندی خوشبندی‌ها بر اساس مفاهیم نزدیکترین همسایه با استفاده از معیار پراکندگی (که در بخش قبل شرح داده شده است) انجام می‌شود که در انجام این کار، ابتدا زوج فاصله‌ی میان مؤلفه‌های خوشبندی را محاسبه می‌کنیم و سپس بر اساس تعاریف بخش قبل نزدیکترین همسایه را حذف می‌کنیم. این فرآیند آن قدر تکرار می‌شود تا تمام خوشبندی‌های باقی‌مانده یا انتخاب شوند یا حذف شوند. در این روش تعداد خوشبندی‌هایی که قرار

¹²⁸ Moderate

است در فرآیند بالا انتخاب شوند از قبل تعریف می‌شوند. مطابق این تعاریف روش پیشنهادی ژیا را به صورت زیر تعریف می‌کنیم: [86]

فرض کنید که M تعداد خوشه‌بندی‌های اصلی باشد و مجموعه‌ی اصلی خوشه‌بندی به صورت $O = \{C_i, i = 1, \dots, M\}$ و فاصله‌ی بین خوشه‌های C_i و C_j به صورت $Sim(C_i, C_j)$ تعریف شوند که در آن برای محاسبه معیار Sim از تعاریف بخش قبل استفاده می‌کنیم و فاصله‌ی بین خوشه‌بندی C_i و نزدیک‌ترین همسایه آن در زیرمجموعه کاهش‌یافته SUB را با r_i^1 نشان می‌دهیم. شبه کد شکل زیر مراحل اجرای روش پیشنهادی ژیا را نشان می‌دهد: [86]

Step 1: Initialize the reduced subset SUB to the original set O ;

Step 2: For each clustering $C_i \in SUB$, compute r_i^1

Step 3: Find clustering C_i for which r_i^1 is the minimum.

Retain it in SUB and discard its nearest neighbors (Note: C_i denotes the clustering for which removing its nearest neighbors will cause minimum error among all clusterings in C_i).

Step 4: Repeat steps 2 and 3 until the predefined number of clusterings is achieved.

Step 5: Return to SUB as the reduced clustering set.

After the selection process, we apply a consensus function to this subset and get a consensus partition.

شکل ۲-۲۹. الگوریتم خوشه‌بندی ترکیبی طیفی مبتنی بر انتخاب بر اساس شباهت [86]

در شکل بالا ابتدا زیرمجموعه کاهش‌یافته SUB ایجاد می‌شود. سپس برای هر خوشه‌بندی C_i مقدار r_i^1 محاسبه می‌شود. آنگاه خوشه‌بندی C_i به نحوی انتخاب می‌شود که مقدار r_i^1 آن کمینه باشد. دو فرآیند محاسبه r_i^1 و پیدا کردن C_i آن قدر تکرار می‌شود تا مطابق با تعداد از قبل تعریف‌شده نتیجه‌ی خوشه‌بندی به دست آوریم. در انتها با استفاده از یک تابع توافقی خوشه‌بندی‌ها را باهم ترکیب کرده و نتیجه نهایی را به دست می‌آوریم. [86].

۴-۴-۲. خوشبندی ترکیبی انتخابی لی مین

این الگوریتم توسط لی مین و همکاران ارائه شده است که ما آن را با عنوان خوشبندی ترکیبی انتخابی^{۱۲۹} لی مین می‌شناسیم. در این الگوریتم ابتدا به این موضوع اشاره می‌شود که چون معمولاً در روش‌های مبتنی بر انتخاب تمامی مقایسه‌ها بر اساس یک افزار مرجع صورت می‌گیرد، کیفیت این افزار بر روی نتایج نهایی تأثیر به سزایی دارد. از این روی در این روش دو رویکرد پیشنهاد می‌شود: ابتدا برای انتخاب بهترین افزار مرجع بر اساس ارزیابی اعتبار خوشبندی راه حلی پیشنهاد می‌شود و سپس یک راهکار انتخاب خوشی جدید با استفاده از وزن اعضا پیشنهاد می‌شود. در ادامه روش پیشنهادی لی مین و همکاران را شرح می‌دهیم:

فرض کنید $\{P_1, P_2, \dots, P_H\}$ مجموعه‌ی N تایی از نقاط داده و $X = \{x_1, x_2, \dots, x_N\}$ مجموعه‌ی H تایی افزارهای خوشبندی بر روی این مجموعه داده می‌باشد. هر افزار P_i شامل مجموعه خوش‌های $C_i^1, C_i^2, \dots, C_i^{k_i}$ می‌باشد که k_i تعداد خوش‌های افزار p_i است و رابطه $X = \bigcup_{j=1}^k C_i^j$ همواره برقرار می‌باشد. هدف این روش انتخاب زیرمجموعه‌ی P' بر اساس راهکار انتخاب خوش پیشنهادی آن (که در ادامه بررسی خواهیم کرد) از مجموعه P و ساخت نتیجه نهایی C^f مطابق با زیرمجموعه‌ی P' با استفاده از تابع توافقی می‌باشد [85].

۱-۴-۴-۲. انتخاب افزار مرجع در روش لی مین

همان طور که پیش‌تر ذکر شد انتخاب افزار مرجع در خوشبندی مبتنی بر انتخاب مهم می‌باشد و تأثیر به سزایی در کیفیت نتیجه نهایی دارد. هر چه افزار مرجع باکیفیت‌تر باشد می‌تواند تأثیر کیفیت پایین نتایج تولیدشده را از بین ببرد. از این روی باکیفیت‌ترین افزار باید به عنوان افزار مرجع انتخاب شود. معمولاً بهترین خوشبندی را به صورتی که اعضای داخل خوش نسبت به هم شبیه‌تر و نسبت به اعضای سایر خوش‌ها متفاوت‌تر باشند تعریف می‌کنیم. در روش پیشنهادی لی مین از غلظت^{۱۳۰} و

¹²⁹ Selective Clustering Ensemble

¹³⁰ Compactness

جدایی^{۱۳۱} به عنوان معیار کیفیت استفاده شده است که ما آن را شرح می‌دهیم. ابتدا میانگین مجموعه داده‌ی X را به صورت $\mu = \frac{1}{N} \sum_{i=1}^N x_i$ ^{۱۳۲} و وردایی^{۱۳۳} مجموعه داده X را به صورت x_j محاسبه می‌کنیم که در آن $d(x_i, x_j)$ فاصله‌ی دو داده‌ی x_i و x_j می‌باشد. مطابق تعاریف بالا غلظت افزار P_i به صورت رابطه زیر تعریف می‌شود [85].

$$Cmp = \frac{\sum_{j=1}^{k_i} MSE(C_i^j)}{MSE(X)} \quad (59-2)$$

این الگوریتم برای بر اساس توزیع گوسی جدایی را مطابق رابطه زیر تعریف می‌کند:

$$Sep = \frac{1}{k_i(k_i - 1)} \sum_{i=1}^c \sum_{j=1, j \neq i}^c \exp \left[-\frac{d^2(\overline{x}_{c_i^n}, \overline{x}_{c_i^m})}{2\sigma^2} \right] \quad (60-2)$$

در رابطه بالا σ ثابت گوسی (بر اساس $1 = 2\sigma^2$ تنظیم می‌شود) و پارامترهای $\overline{x}_{c_i^n}$ و $\overline{x}_{c_i^m}$ به ترتیب مراکز C_i^n و C_i^m می‌باشد. برای اعتبارسنجی کیفیت خوشها روش لی مین مطابق رابطه زیر دو معیار غلظت و جدایی را در یکتابع مشترک ادغام می‌کند: [85]

$$OBJ = \alpha \times Cmp + (1 - \alpha) \times Sep \quad (61-2)$$

در این رابطه پارامتر α برای کنترل تأثیرات دو معیار غلظت و جدایی است که در این روش 0.5 در نظر گرفته شده است تا تأثیرات دو معیار برابر باشد. در ادامه این روش مقدار OBJ را برای تمامی $P = \{P_1, P_2, \dots, P_H\}$ محاسبه کرده و افزایی که مقدار OBJ آن کمینه باشد را به عنوان افزار مرجع در نظر می‌گیرد و آن را P^* نام‌گذاری می‌کند [85].

¹³¹ Separation

¹³² Variance

۲-۴-۴-۲. راهکار انتخاب خوش در روش لی مین

منظور از راهکار انتخاب خوش، گزینش بهترین افزایهای خوشبندی از میان تمامی افزایهای تولید شده می‌باشد. انتخاب راهکار مناسب موجب می‌شود که در افزایهای انتخاب شده ویژگی‌های ضمنی مجموعه داده منعکس شود و عملکرد خوشبندی را بهبود یابد. روش لی مین یک راهکار بر اساس معیار کیفیت و پراکندگی برای انتخاب خوش پیشنهاد می‌کند که ما آن را در ادامه بررسی خواهیم کرد [85].

در روش‌های بدون ناظر، مسئله خوشبندی بدون برچسب عمل می‌کند. بنابراین تناظر روشی بین نتایج فراهم شده با استفاده از خوشبندی مختلف نیست. برای مثال نتایج دو خوشبندی $P_1 = \{1, 1, 2, 2, 3, 3\}$ و $P_2 = \{3, 3, 1, 1, 1, 2, 2\}$ مشابه در نظر گرفته می‌شوند. برای حل این مشکل این روش یک ماتریس اتصال¹³³ را تعریف می‌کند. در این تعریف مجموعه افزایهای

$P = \{P_1, P_2, \dots, P_H\}$

$$M^i(j, k) = \begin{cases} 1 & \text{if } P_j^i = P_k^i \quad (1 \leq j \leq N, 1 \leq k \leq N) \\ 0 & \text{otherwise} \end{cases} \quad (62-2)$$

مطابق رابطه بالا شکل زیر ماتریس اتصال مثال ذکر شده (برای افزایهای P_1 و P_2) را نشان می‌دهد:

$$M^1 = M^2 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

شکل ۲-۳۰. مثالی از ماتریس اتصال [85].

حال فرض کنید برای مجموعه افزایهای $(1 \leq i \leq H)$ و افزای مرجع P^* به ترتیب ماتریس‌های اتصال M_i و M^* موجود می‌باشند. شباهت خوشبندی‌ها به صورت رابطه زیر تعریف می‌شود:

¹³³ Connectivity Matrix

$$N_{p_i}^{(q)} = \frac{\sum_{i=1}^H \|M_i, M^*\|}{H} \quad (63-2)$$

در رابطه بالا $\|M_i, M^*\|$ توافق^{۱۳۴} (اشاره به معیار توافقی برای ارزیابی) بین دو ماتریس است که رویکردهای متنوعی برای محاسبه آن وجود دارد. روش پیشنهادی لی مین از معیار NMI برای محاسبه $\|M_i, M^*\|$ استفاده می‌کند. مطابق با رابطه بالا می‌توان نتیجه گرفت که شبیه‌ترین P_i به P^* و باکیفیت‌ترین P^* دارای بالاترین مقدار $N_{p_i}^{(q)}$ می‌باشد. به طور مشابه، پراکندگی خوشبندی به صورت رابطه زیر تعریف می‌شود: [85]

$$N_{p_i}^{(d)} = 1 - \frac{\sum_{i=1, j \neq i}^H \|M_i, M_j\|}{H(H-1)} \quad (64-2)$$

بدیهی است که P_i با بیشترین پراکندگی دارای بالاترین مقدار $N_{p_i}^{(d)}$ می‌باشد. در روش لی مین استفاده همزمان معیار کیفیت و پراکندگی به عنوان یک راهکار مناسب جهت انتخاب خوش معرفی می‌شود. از آنجایی که گاهی اوقات این دو عامل متناقض در نظر گرفته می‌شوند، این روش یک تناظر^{۱۳۵} بین این دو معیار در افزایش‌های برقرار می‌کند. از این روی رابطه زیر توسط لی مین برای ترکیب کیفیت و پراکندگی پیشنهاد می‌شود: [85]

$$CF(P_i) = \lambda N_{p_i}^{(a)} + (1-\lambda) N_{p_i}^{(d)} \quad (65-2)$$

در رابطه بالا پارامتر λ برای متعادل کردن معیارهای پراکندگی و کیفیت استفاده می‌شود که در این روش پیشنهادی ۰.۵ در نظر گرفته شده است. با محاسبه رابطه (۷-۱) برای تمامی افزایش‌ها در خوشبندی‌ها مطابق روش لی مین می‌توان یک ارزیابی از خوش‌های داشت که بر اساس آن K تا از بهترین افزایش‌ها ($1 \leq K \leq H$) جهت ساخت کمیته^{۱۳۶} ترکیب انتخاب می‌شود. قابل ذکر است که اگر K برابر با H باشد تمامی افزایش‌ها در ترکیب نهایی شرکت می‌کنند و اگر K برابر با یک باشد فقط یک افزایش نتیجه نهایی را می‌سازد و در صورتی که $H \leq K \leq 1$ باشد متناسب با تعادل بین دقت و پراکندگی افزایش‌ها انتخاب می‌شوند [85].

¹³⁴ Agreement

¹³⁵ Trade-off

¹³⁶ Committee

۳-۴-۴-۲. چهارچوب الگوریتم خوشبندی انتخابی لی مین

از آنجایی که اثر انتخاب افزارها در نتیجه خوشبندی نهایی متفاوت است در این روش یک وزن برای کنترل الگوریتم پیشنهاد می‌شود. برای هر افزار انتخاب شده، میانگین مقادیر تابع CF مطابق با رابطه ۷-۱ به عنوان وزن آن افزار به صورت زیر انتخاب می‌شود: [85]

$$\varphi_{P_i} = \frac{\sum_{i=1, \dots, K} CF(P_i)}{K} \quad (66-2)$$

افرازی که بیشترین تأثیر در ساخت نتایج نهایی را دارد دارای مقدار بیشینه φ_{P_i} خواهد بود. بنابراین رابطه وزن به صورت زیر تعریف می‌شود:

$$W_{P_i} = \varphi_{P_i} \times \frac{1}{\chi} (i = 1, \dots, K) \quad (67-2)$$

که χ برای نرمال سازی وزن استفاده می‌شود. از این روی $W_{P_i} > 0$ و $1 < W_{P_i} < 0$ می‌باشد. در این روش بعد از اختصاص وزن به تمامی افزارهای انتخاب شده، آنها را با استفاده از تابع توافقی $CSPA$ ادغام می‌کنیم. شکل زیر شبه کد الگوریتم پیشنهادی لی مین را نشان می‌دهد [85].

Step1: Using some clustering algorithm (runs H times) to get H clustering partitions.

Step2: Calculating objective function OBJ of all available clustering partitions and choosing the minimum (that is P^*) as the reference partition.

Step3: Calculating criterion function $CF(P_i)$ and choosing K best clustering partitions.

Step4: Calculating the weight of K best clustering partitions.

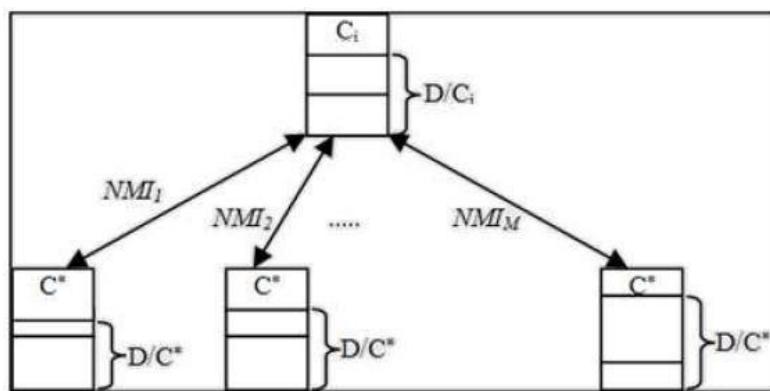
Step5: Using the CSPA consensus function to produce the final result.

شکل ۳۱-۲. شبه کد خوشبندی ترکیبی انتخابی لی مین [85]

۴-۲ خوشبندی بر اساس معیار **MAX** با استفاده از مجموعه‌ای از خوشبندی‌های یک افزار

این روش توسط علیزاده و همکاران ارائه شده است که در آن به جای ارزیابی افزار به صورت کامل (استفاده از تمامی خوشبندی‌های افزار) به ارزیابی خوشبندی‌ها به صورت مجزا می‌پردازد و سپس با استفاده از مجموعه‌ای از آن خوشبندی‌ها جواب نهایی را تشکیل می‌دهد. در این روش برای حل مشکل NMI راهکاری با عنوان روش MAX ارائه می‌شود [9]. در ادامه ابتدا به بررسی راهکارهای پیشنهادی در این روش می‌پردازیم.

۴-۳ راهکار ارزیابی خوشبندی **MAX**



شکل ۴-۲. روش ارزیابی خوشبندی یک افزار در روش MAX

شکل ۴-۲ روش ارزیابی خوشبندی C_i را با خوشبندی‌ای از سایر افزارهای ساخته شده برای ترکیب نشان می‌دهد. در این شکل برای حل مشکل تقارن NMI بقیه خوشبندی‌های افزاری که در آن C_i وجود دارد را با عنوان D/C_i می‌شناسیم و خوشبندی‌ای که قرار است با خوشبندی C_i مقایسه شود را C^* و سایر خوشبندی‌های افزار مقایسه شونده را D/C^* می‌نامیم. علیزاده و همکاران در [9] اثبات کردند که در این حالت مشکل تقارن در NMI حل می‌شود. همچنین در این روش معیاری تحت عنوان پایداری برای ارزیابی خوشبندی‌ها ارائه شده است.

$$Stability(C_i) = \sum_M NMI_i \quad (48-2)$$

در رابطه بالا پارامتر M تعداد خوشبندی‌های افزار مرجع را نشان می‌دهد و اطلاعات متقابل نرمال شده مطابق رابطه ۶۹-۲ به ازای C_i و سایر خوشبندی‌های افزار مرجع محاسبه می‌شود.

$$NMI(P^a, P^b) = \frac{-2 \sum_{i=1}^{k_a} \sum_{j=1}^{k_b} n_{ij}^{ab} \log \left(\frac{n_{ij}^{ab} \times n}{n_i^a \times n_j^b} \right)}{\sum_{i=1}^{k_a} n_i^a \log \left(\frac{n_i^a}{n} \right) + \sum_{j=1}^{k_b} n_j^b \log \left(\frac{n_j^b}{n} \right)} \quad (69-2)$$

در رابطه بالا n تعداد کل نمونه‌ها و n_{ij}^{ab} تعداد الگوهای مشترک بین $C_i^a \in P^a$ و $C_j^b \in P^b$ و n_i^a تعداد الگوهای خوشه‌ی i از افزار a و n_j^b تعداد الگوهای خوشه‌ی j از افزار b می‌باشد.

۲-۵-۴-۲. روش انباشت مدارک توسعه‌یافته

از آن جایی که در روش MAX تمامی خوشه‌های یک افزار در ساخت نتیجه نهایی وجود ندارد نمی‌توان از روش کلاسیک برای ساخت ماتریس همبستگی استفاده کرد، از این روی علیزاده و همکاران روش انباشت مدارک توسعه‌یافته (EEAC¹³⁷) را معرفی کردند [8, 9, 67] که برای ساخت ماتریس همبستگی از رابطه (70-۲) استفاده می‌کند.

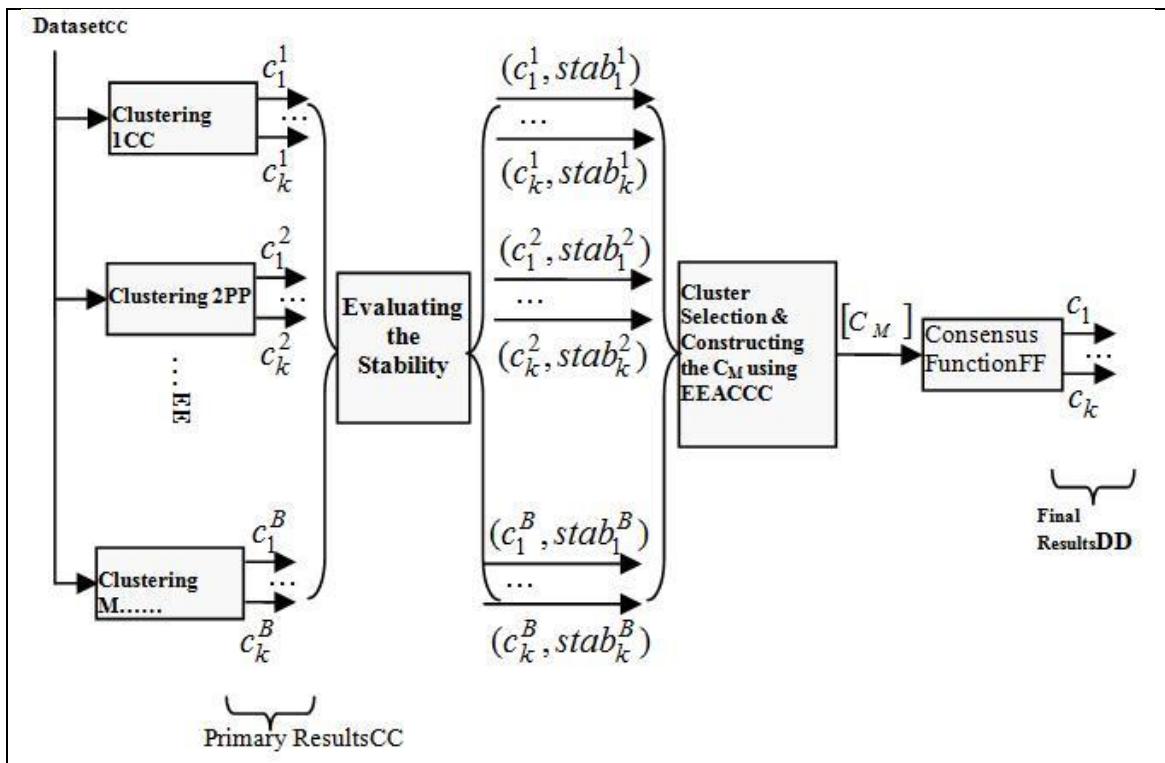
$$C(i, j) = \frac{n_{i,j}}{\max(n_i, n_j)} \quad (70-2)$$

در این رابطه n تعداد دفعاتی است که نمونه‌ی در خوشه‌های انتخاب شده ظاهر شده است. به طور مشابه n نیز، تعداد دفعاتی است که نمونه‌ی زدرا خوشه‌های انتخاب شده ظاهر شده است. همچنین $n_{i,j}$ تعداد دفعاتی است که جفت نمونه‌های i و j باهم در یک خوشه از خوشه‌های انتخاب شده ظاهر شده‌اند. بدیهی است که با در نظر گرفتن تعداد خوشه‌های ثابت در خوشبندی‌های اولیه همواره n و n کمتر از تعداد کل افزارهای اولیه و همچنین، تعداد کل خوشه‌های ممکن می‌باشد.

۲-۶-۴. خوشبندی بر اساس معیار APMM با استفاده از مجموعه‌ای از خوشه‌های یک افزار

این روش توسط علیزاده و همکاران بر اساس معیار APMM (رابطه ۲۹-۲) ارائه شده است که در آن همانند روش MAX، تعدادی از خوشه‌های یک افزار در تهیه نتیجه نهایی به کار گرفته می‌شوند. در این روش مقدار پایداری خوشه بر اساس رابطه ۳۰-۲ ارزیابی شده و با استفاده از روش انباشت مدارک اصلاح توسعه‌یافته (رابطه ۷۰-۲) نتیجه‌ی نهایی را می‌سازیم [8, 9, 67].

¹³⁷ Extended EAC

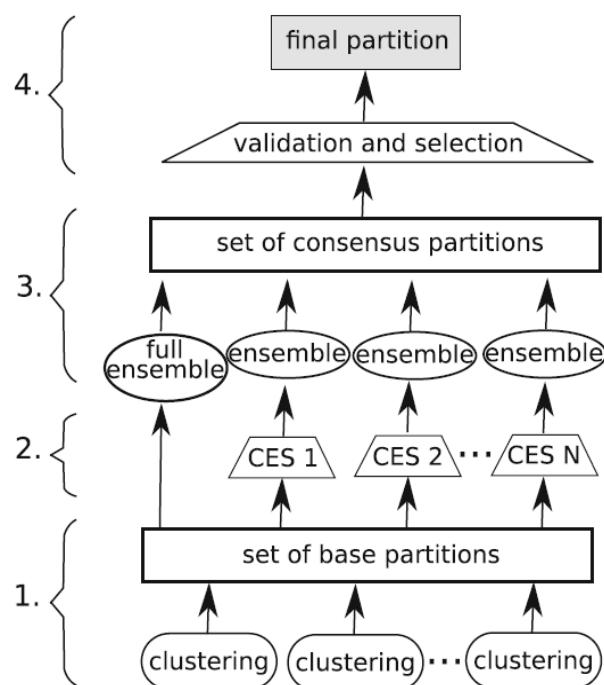


شکل ۲-۳۳. چهارچوب خوشبندی ترکیبی مبتنی بر انتخاب با استفاده از مجموعه‌ای از خوشبندی‌های یک افزار [8, 9]

شکل ۲-۳۳-۲ چهارچوب خوشبندی ترکیبی مبتنی بر انتخاب با استفاده از مجموعه‌ای از خوشبندی‌های یک افزار را نشان می‌دهد. در این روش ابتدا الگوریتم‌های خوشبندی پایه نتایج اولیه را تولید می‌کنند و سپس از ارزیابی پایداری (با استفاده از معیار MAX و یا APMM) خوشبندی‌های افزارهای تولید شده تعدادی از آنها را انتخاب می‌کنیم. قابل ذکر است که این تعداد در ابتدا توسط کاربر برای الگوریتم تعریف می‌شود. علیزاده و همکاران عمل انتخاب خوشبندی ارزیابی شده برای شرکت در ترکیب نهایی را آستانه‌گیری می‌نامند. مطابق شکل ۲-۳۳-۲ پس از تشکیل مجموعه نهایی ترکیب با استفاده از روش انباشت مدارک توسعه یافته اقدام به تولید نتیجه نهایی می‌کنیم. این تحقیق به توسعه یک معیار ارزیابی افزار مبتنی بر معیار AMPP می‌پردازد چون اولاً این معیار مشکل تقارن NMI را ندارد و در ثانی کارایی بالای آن اثبات شده است [1, 8, 67].

۲-۵. روش بهترین افزار توافقی اعتبارسنجی شده

روش بهترین افزار توافقی اعتبارسنجی شده^{۱۳۸} یا همان BVCP که اخیراً توسط نالدی و همکاران ارائه شده است به این موضوع اشاره می‌کند که معیارهای ارائه شده (شاخص‌های نسبی اعتبارسنجی خوشبندی^{۱۳۹}) جهت ارزیابی و انتخاب افزار و یا خوشبندی در روش‌های خوشبندی ترکیبی مبتنی بر انتخاب هر کدام بر روی یک سری مجموعه داده بهتر عمل می‌کنند. از این روی مطابق شکل زیر در این روش پیشنهاد می‌شود تا از نتایج ترکیب کامل به همراه چندین ترکیب مبتنی بر انتخاب (CES) برای ساخت نتیجه‌ی نهایی استفاده شود [87].



شکل ۲-۳۴. چهارچوب روش بهترین افزار توافقی اعتبارسنجی شده [87]

در این روش مطابق شکل بالا ابتدا توسط الگوریتم‌های خوشبندی پایه نتایج اولیه را ایجاد می‌کنیم. افزارهای به دست آمده در این بخش توسط الگوریتم خوشبندی ترکیب کامل و چند الگوریتم خوشبندی ترکیبی مورد استفاده قرار می‌گیرند. در این روش پس از ساخت ماتریس‌های همبستگی با استفاده از روش‌های ترکیبی آنها را در مجموعه‌ای جهت ارزیابی و انتخاب جمع‌آوری می‌کنیم.

¹³⁸ The Best Validated Consensus Partition method

¹³⁹ Relative clustering validity index(es)

پس از ارزیابی نتایج ماتریس‌های همبستگی بهترین افزار به عنوان جواب نهایی انتخاب می‌شود. در این روش شاخص‌های ^{۱۴۰}SS، ^{۱۴۱}ASS، ^{۱۴۲}VRC، ^{۱۴۳}PBM و ^{۱۴۳}DB Dunn جهت ارزیابی ماتریس‌های همبستگی معرفی شده‌اند. قابل ذکر است که مسائلی همچون پیچیدگی زمانی، روش انتخاب شاخص‌ها و الگوریتم‌های خوشبندی ترکیبی و تأثیرات توابع توافقی مختلف بر روی نتیجه نهایی در این روش با ابهامات زیادی رو به رو است. از آن جایی که تمرکز این تحقیق بر روی ارائه‌ی روش‌های خوشبندی ترکیبی مبتنی بر انتخاب می‌باشد و نه ارزیابی و ترکیب نتایج آن‌ها بنابراین تشریح روش کار این الگوریتم خارج از محدوده‌ی کاری این تحقیق می‌باشد.

۶-۲. استفاده از نظریه خرد جمعی در علوم رایانه

استفاده از نظریه‌های مطرح شده در سایر علوم از نظری ریاضیات، اقتصاد، مدیریت و غیره در علوم رایانه امری بعید و دور از انتظار نیست. کارایی نظریه خرد جمعی که اولین بار توسط سورویکی در سال ۲۰۰۴ در کتابی با همین عنوان مطرح شده است، توسط مثال‌های فراوانی از قضایای اجتماعی، اقتصادی و غیره اثبات شده است. از این روی در سال‌های اخیر از نظریه خرد جمعی در چند تحقیق در حوزه‌ی علوم رایانه استفاده شده است که ما به برخی از آن‌ها به صورت مختصر اشاره می‌کنیم. استیورز^{۱۴۴} و همکاران از خرد جمعی برای باز جمع‌آوری اطلاعات مرتب شده استفاده کرده‌اند [92] و میلر^{۱۴۵} و همکاران یک رویکرد جدید برای رتبه‌بندی مسائل مرتب‌شده ارائه داده‌اند [91]. ولی ندر^{۱۴۶} و همکاران از این نظریه برای تخمین مقادیر اساسی (نظیر کلاس^{۱۴۷}) در طبقه‌بندی^{۱۴۸}) برای هر تصویر در محیط نویزی استفاده کرده‌اند که آن را خرد جمعی چندبعدی^{۱۴۹} نامیده‌اند [89]. ویلیام و همکاران راهکاری با عنوان طبقه‌بندی معدن زیر آب^{۱۵۰} با استفاده از برچسب‌های بدون کیفیت بر اساس نظریه خرد جمعی ارائه داده‌اند [90]. بی و همکاران بر اساس این

¹⁴⁰ Simplified Silhouette

¹⁴¹ Alternative Simplified Silhouette

¹⁴² Calinski-Harabasz

¹⁴³ Davies-Bouldin

¹⁴⁴ Steyvers

¹⁴⁵ Miller

¹⁴⁶ Welinder

¹⁴⁷ Class

¹⁴⁸ Classification

¹⁴⁹ Multidimensional wisdom of crowds

¹⁵⁰ Underwater mine classification

نظریه روشی برای مسئله درخت پوشای حداقل^{۱۵۱} ارائه داده‌اند [88]. بیکر و همکاران با استفاده از نظریه خرد جمعی در محیط‌های مدل‌سازی یک روش برای طبقه‌بندی ترکیبی ارائه داده‌اند [10]. لازم به ذکر است که این تحقیق در نگاشت شرایط چهارگانه خرد جمعی به خوشه‌بندی ترکیبی مبنی بر انتخاب از مفاهیم و ایده‌های مطرح شده در [10] استفاده کرده است که در بخش سوم به بررسی آن‌ها می‌پردازیم.

¹⁵¹ Minimum spanning tree

فصل سوم

روش تحقیق

۳. روش تحقیق

۱-۱. مقدمه

در این فصل ابتدا به بررسی نظریه خرد جمعی بر اساس تعاریف کتاب سورویکی می‌پردازیم. در این راستای ابتدا شرایط چهارگانه جامعه خردمند را بررسی می‌کنیم و سپس به بررسی استشناها در این نظریه خواهیم پرداخت. پس از تشریح نظریه خرد جمعی، روش پیشنهادی اول این تحقیق را شرح می‌دهیم. این روش را ما با عنوان "خوشبندی خردمند با استفاده از آستانه‌گیری" می‌شناسیم که در آن مطابق با ادبیات مطرح شده در خوشبندی ترکیبی، ابتدا به بیان چهارچوب کلی آن پرداخته و سپس برای شرایط چهارگانه خرد جمعی تعریفی مناسب و جدید ارائه می‌دهیم. آنگاه بر اساس این تعاریف الگوریتم پیشنهادی روش اول را بیان می‌کنیم. در این روش پس از تولید نتایج اولیه با استفاده از چهارچوب پیشنهادی غیر مرکز به آستانه‌گیری از نتایج به دست آمده از ارزیابی درجه استقلال الگوریتم‌ها و پراکندگی نتایج اولیه می‌پردازیم و در پایان بر اساس نتایج انتخاب شده (افرازهای خردمند) نتیجه نهایی را تولید می‌کنیم. در این روش دو الگوریتم پایه غیر هم نام کاملاً مستقل و درجه استقلال الگوریتم‌های هم نام بر اساس پارامترهای اساسی آنها محاسبه می‌شود.

در ادامه، روش پیشنهادی دوم این تحقیق بیان می‌شود. این روش که ما آن را با عنوان "خوشبندی خردمند مبتنی بر گراف استقلال الگوریتم" می‌شناسیم به بهبود دو بخش از روش اول می‌پردازد. این روش در ابتدا این ایده را بررسی می‌کند که الگوریتم‌های غیر هم نام کاملاً مستقل نیستند. در این راستای برای محاسبه درجه استقلال دو الگوریتم با استفاده از ایده تبدیل کد به گراف در تست نرم‌افزار به ارائه روشی با عنوان مدل‌سازی گراف استقلال الگوریتم می‌پردازیم. با مقایسه گراف‌های به دست آمده (درجه استقلال) در این روش می‌توان یک وزن برای احتمال صحت جواب‌های به دست آمده پیشنهاد داد. از این روی در این روش به جای ارزیابی و آستانه‌گیری از استقلال الگوریتم‌ها، آنها را به عنوان وزنی برای ترکیب نتایج در نظر می‌گیریم که این کار نیاز به تعیین آستانه برای استقلال را از بین می‌برد. در روش دوم پیشنهادی این تحقیق رابطه‌ای جدید بر اساس

رابطه ۲-۵۶ برای ترکیب نتایج اولیه به صورت وزن دار با عنوان روش انباشت مدارک وزن دار یا WEAC^{۱۵۲} معرفی می شود.

۲-۳. نظریه خرد جمعی

"فرانسیس گالتون" فیلسوف و دانشمند علم آمار از انگلستان بود که مفاهیم اصلی انحراف استاندارد و همبستگی را معرفی کرد. یک روز که او از نمایشگاه دام بازدید می کرد، به جایی رسید که در آن مسابقه‌ای ترتیب داده شده بود. یک گاو نر فربه انتخاب شده و در معرض دید عموم قرار گرفته بود. هر کس که تمایل شرکت در مسابقه را داشت باید شش پنس می پرداخت و ورقه‌ای مهرشده را تحویل می گرفت. در آن ورقه باید تخمین خود را از وزن گاو نر می نوشت. نزدیک‌ترین تخمین به واقعیت برنده مسابقه بود و جوازی به صاحب آن تعلق می گرفت. مجموعاً ۷۸۷ نفر در مسابقه شرکت کردند تا شанс خود را بیازمایند. افراد از همه تیپ و طبقه‌ای آمده بودند. از قصاب گرفته که قاعده‌تاً باید بهترین و نزدیک‌ترین نظر را به واقعیت می داد تا کشاورز و مردم عامی بی تخصص. گالتون می خواست دریابد عقل جمعی مردم پلیموت چگونه قضاوت کرده است. بدون شک تصور او این بود که عدد مزبور فرسنگ‌ها از عدد واقعی فاصله خواهد داشت چرا که از دید وی افراد کم‌هوش و عقب‌مانده در آن جمع اکثریت قاطع را تشکیل می دادند. برخلاف نظر گالتون، متوسط نظرات جمعیت این بود که گاو نر ۱۱۹۷ پوند وزن دارد و وزن واقعی گاو که در روز مسابقه وزن‌کشی شد ۱۱۹۸ پوند بود. گالتون اشتباه می کرد. نظر جمع تقریباً به طور کامل با واقعیت تطابق داشت گالتون در مقاله‌ای که در مجله علمی "طبیعت" منتشر نمود نوشت نتایج نشان می دهد که قضاوت‌های جمعی و دموکراتیک از اعتبار بیشتری نسبت به آنچه که من انتظار داشتم برخوردارند [44]. نظریه هیئت‌منصفه که اولین بار توسط کندورست بیان شد نیز این نتیجه گالتون را تأیید می کند. این نظریه در علوم سیاسی، احتمال نسبی درستی نظر گروهی از افراد (رأی اکثریت) را بررسی می کند. نظریه خرد جمعی که اولین بار توسط سورویکی در کتابی با همان عنوان انتشار یافته است، تأیید می کند که یک جمع می تواند مسئله را بهتر از اکثر اعضای گروه حل کند. مطابق تعریف این کتاب، یک جمعیت به هر گروهی از افراد اطلاق می شود که می توانند به طور جمعی تصمیمی بگیرند یا مسئله‌ای را حل کنند [55].

¹⁵² Weighted Evidence Accumulation Clustering

۱-۲-۳. شرایط جامعه خردمند

خرد جمعی روشنی نوین برای تصمیم‌گیری‌های اجتماعی می‌باشد. کارایی این روش نه تنها در نظریه بلکه در عمل نیز در مسایل مختلف اثبات شده است که پیش‌تر به آن اشاره شده است. این روش تأیید می‌کند که یک جمع می‌تواند مسئله را بهتر از اکثر اعضای گروه حل کند مطابق تعریف این خرد جمعی، یک جمعیت به هر گروهی از افراد اطلاق می‌شود که می‌توانند به طور جمعی تصمیمی بگیرند یا مسئله‌ای را حل کنند. مطابق تحقیقات مک‌کی، همه جمعیت‌ها (گروه‌ها) خردمند نیستند. یک مثال روشن از این قضیه بازار سهام است که جمعیت به سمت حباب بازار هدایت می‌شود. بنابراین ابتدا باید فهمید که تحت چه شرایطی خرد جمعی می‌تواند اثرگذار باشد [44]. از این روی در این روش چهارچوبی جهت تعریف جامعه خردمند ارائه شده است. سورویکی چهار شرط اساسی زیر را برای تمایز جمعیت خردمند از یک جمعیت غیر عاقل پیشنهاد می‌دهد: [55]

۱- نوع (پراکندگی^{۱۵۳}) آراء

۲- استقلال^{۱۵۴} آراء

۳- عدم تمرکز^{۱۵۵} آراء

۴- روش ترکیب^{۱۵۶} مناسب

۱-۱-۲-۳. تعریف معیار پراکندگی

در خرد جمعی معیار نوع یا پراکندگی به صورت زیر تعریف می‌شود:

"هر فرد باید به طور جداگانه اطلاعی از موضوع مورد نظر داشته باشد حتی اگر اطلاعات مزبور غلط و مخدوش باشد." [55]

¹⁵³ Diversity

¹⁵⁴ Independence

¹⁵⁵ Decentralization

¹⁵⁶ Aggregation method

یکی از دلایلی که سورویکی در خصوص چرایی کارکرد نظریه خرد جمعی مطرح می‌کند این است که نظر هر فرد دو عنصر را در درون خود دارد اطلاعات صحیح و غلط. اطلاعات صحیح (از آن رو که صحیح‌اند) هم جهت‌اند و بر روی یکدیگر انباشته می‌شوند اما خطاهای در جهات مختلف و غیر همسو عمل می‌کنند لذا تمایل به حذف یکدیگر دارند. نتیجه این می‌شود که پس از جمع نظرات آنچه که می‌ماند اطلاعات صحیح است. از این روی معیار پراکنده‌گی یکی از مهم‌ترین اصل‌ها در خرد جمعی است زیرا به طور مستقیم بر روی میزان هم جهت سازی آرای تأثیر دارد [55].

۲-۱-۲-۳. تعریف معیار استقلال

استقلال یکی دیگر از اصل‌های خرد جمعی است. در صورت وابستگی افراد به گروه یا فرد خاصی اصل هم جهتی در آرای از بین می‌رود. به عبارت دیگر در صورت نبود استقلال مقدار انحراف معیار آرای جامعه آماری ما واقعی نخواهد بود و نظریه خرد جمعی در این‌گونه جوامع که ما آن را در با عنوان جامعه‌های دیکتاتوری می‌شناسیم درست عمل نمی‌کند. در خرد جمعی معیار استقلال به صورت زیر تعریف می‌شود:

"نظر افراد باید به طور مستقل و بدون تأثیر گرفتن از یک فرد یا گروه مشخص شکل گیرد." [55]

۳-۱-۲-۳. تعریف معیار عدم تمرکز

در خرد جمعی معیار عدم تمرکز به صورت زیر تعریف می‌شود:

"افراد باید توانایی شخصی سازی و نتیجه‌گیری مبنی بر دانش محلی خود را داشته باشند." [55]

با توجه به مثال‌هایی که سورویکی در مورد عدم تمرکز در آژانس اطلاعات مرکزی آمریکا (CIA^{۱۵۷}) یا سیستم عامل لینوکس ذکر می‌کند، باید گفت که این معیاری کیفی است. همچنین فاجعه شاتل کلمبیا یکی از مثال‌های مهمی است که در کتاب خرد جمعی در مورد مشکلات بالقوه تمرکز به آن اشاره شده است. سورویکی این مشکل را این‌گونه توجیه می‌کند، به علت بوروکراسی در مدیریت سلسله مراتبی ناسا این فاجعه کاملاً به آگاهی مهندسین سطح پایین (اجرایی) وابسته شده بود (امکان ردیابی آن در سطوح بالاتر وجود نداشت) وی همچنین اشاره می‌کند که چون تمامی مهندسان ناسا به

¹⁵⁷ Central Intelligence Agency

صورت متمرکز برای این پروژه آموزش دیده بودند از این رو هیچ کس مشکل را درک نکرد که این منجر به آن فاجعه شد [55].

۴-۱-۲-۳. روش ترکیب مناسب

یکی دیگر از اصل‌های بسیار مهم در خرد جمعی روش ترکیب مناسب می‌باشد. سورویکی آن را این‌گونه تعریف می‌کند:

"باید مکانیزمی وجود داشته باشد که بتوان توسط آن نظرات افراد را با یکدیگر ترکیب کرده و به یک نظر جمعی تبدیل نمود" [55]

آن چیز که در اینجا بدھی است آن است که روش ترکیب باید طوری انتخاب شود که مناسب با داده‌های ورودی (رأی افراد) باشد و خروجی مناسب را تولید کند (یک نظر واحد و کامل) و در ادغام داده‌های ورودی کمترین خطأ را داشته باشد (به طور ایده‌آل صفر) و پاسخگوی انواع مسائل خرد جمعی باشد. از این روی نمی‌توان یک روش واحد برای ترکیب نتایج اولیه تمامی مسائل خرد جمعی پیدا کرد و باید مناسب با هر مسئله روشی را اتخاذ کرد برای مثال در مسئله فروشگاه دام که گالتون آن را با ایده ابتدایی خرد جمعی حل کرد مکانیزم ترکیب "میانگین" بود.

۴-۲-۳. اهمیت و رابطه استقلال و پراکندگی در خرد جمعی

در سال‌های نخستین قرن بیستم طبیعی دان آمریکائی "ویلیام بیب"^{۱۵۸} در حین مطالعات خود در جنگل‌های جزایر گویان با منظره عجیبی برخورد کرد. لشگر بزرگی از مورچه‌ها در پیرامون یک دایره بزرگ که محیطی در حدود ۴۰۰ متر داشت بی‌وقفه در حال حرکت بودند. آنان هر ۲/۵ ساعت یکبار به دور این دایره می‌گشتند. این گردش آن قدر ادامه یافت که پس از ۲ روز اکثر آن‌ها جان خود را از دست دادند. آنچه که بیب مشاهده کرده بود بیولوژیست‌های امروزی آن را "دایره آسیاب"^{۱۵۹} می‌نامند. این دایره زمانی شکل می‌گیرد که گروهی از مورچگان از "جمع"^{۱۶۰} خود به دور می‌افتد. وقتی که چنین امری اتفاق می‌افتد آنان از یک قانون ساده پیروی می‌کنند. از مورچه جلوی خود تبعیت کن.

¹⁵⁸ William Beebe

¹⁵⁹ Circular Mill

¹⁶⁰ Colony

این دایره زمانی می‌شکند که به طور تصادفی یکی از مورچه‌ها به دلیلی نامعلوم دایره را ترک می‌کند و مورچه بعدی به دنبال او به راه می‌افتد.

جانسون در کتاب خود بنام ظهور می‌گوید: "کلنی مورچگان معمولاً بسیار خوب کار می‌کند. هیچ کس گروه را ترک نمی‌کند، هیچ کس فرمان نمی‌دهد و هیچ کس اطاعت نمی‌کند. هیچ مورچه‌ای به تنها یک نمی‌داند چه می‌کند و هیچ نوع اطلاعاتی در اختیار ندارد اما جمع آن‌ها غذا را پیدا می‌کند، ذخیره می‌کند، کارهای مربوط به جمع را به بهترین شکل انجام می‌دهد و تولید مثل نیز می‌کند". اما همین اصل تبعیت کورکورانه، باعث مرگ آنان در دایره آسیاب می‌شود. یک مورچه هیچ استقلال رأیی ندارد و به همین دلیل هم زمانی که در دایره مرگ گرفتار می‌آید راه خلاصی به بیرون را نمی‌یابد [36]. انسان‌ها اما به خلاف مورچگان می‌توانند مستقل فکر کرده و مستقل عمل کنند. مفهوم استقلال این است که به طور نسبی و به میزانی فرد قادر است مستقل از جمع عمل نماید. این تفاوت مهم و چشمگیری است که جمع ما را از مورچگان متمایز می‌کند.

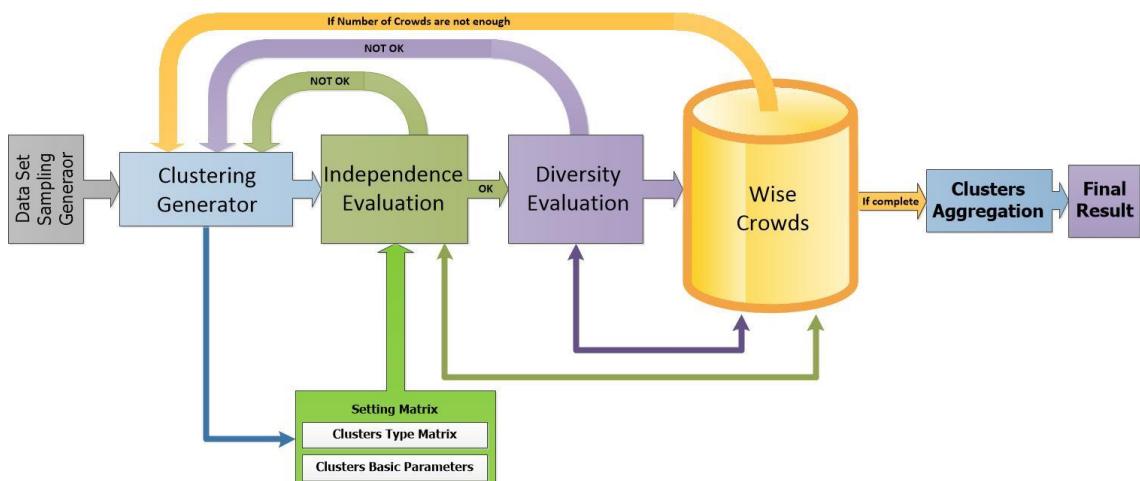
استقلال به دو دلیل از اهمیت بسیاری در ارتقاء هوش جمعی برخوردار است. اول اینکه از تکرار یک نوع خطا جلوگیری می‌کند. خطای یک فرد بر قضاوت یک جمع یک تأثیر خردکننده ندارد، اما اگر همان خطای طور سامانمند در تعداد زیادی از افراد جمع گسترش یابد آن وقت است که رأی جمع را به طور منفی تحت تأثیر قرار می‌دهد. دوم آن که افکار مستقل اطلاعات تازه و متنوع را وارد جمع می‌کند در حالی که اگر افکار مستقل نباشند همان نوع اطلاعات در جمع تکرار می‌شود و چیز تازه‌ای به خرد جمع اضافه نمی‌شود. بنابراین هوشمندترین گروه‌ها آن‌هایی هستند که افراد آن از تنوع بالا و استقلال رأی هرچه بیشتر برخوردار باشند. مفهوم مخالف آن این است که جمعی که افرادش به لحاظ فکری به هم نزدیک و نزدیک‌تر شوند از درجه هوش چندان بالایی برخوردار نیست.

آنچه که ما می‌خواهیم به عنوان یک اصل مهم از آن یاد کنیم این است که هر چقدر افراد یک جمع به یکدیگر نزدیک‌تر باشند و بتوانند با یکدیگر روابط فردی برقرار کنند تصمیم جمع از عقلانیت بیشتر به دور خواهد بود. هر چقدر ما به یکدیگر نزدیک‌تر باشیم باورهایمان به یکدیگر نزدیک شده و امکان تصحیح خطاهایمان کاهش می‌یابد. ممکن است به لحاظ فردی در اثر این هم نشینی خود به هوش و دانش بالاتری دست یابیم اما قطعاً جمع را به بی‌خردی و بلاحت نزدیک می‌کنیم.

۳-۲-۳. استثناءها در خرد جمعی

سه مسئله مجازایی که مشخص شده است که در آنها جمیعت‌ها ممکن است از تک‌تک اعضا هوشمندانه‌تر عمل کنند عبارت‌اند از: الف) مسئله سوزن در انبار کاه که بعضی از افراد جمیعت ممکن است جواب را بدانند در حالی که خیلی‌ها نمی‌دانند. ب) مسئله تخمین حالت که بعضی افراد جمیعت ممکن است با خوش‌شانسی جواب دقیق را بدهنند (در حالی که خودشان از قبل از میزان دقت جوابشان آگاه نباشند)، اما گروه این‌طور نباشد. ج) مسئله پیشگویی که جواب هنوز باید کشف (آشکار) شود [49, 53]. برای مسئله پیشگویی، جواب کشف نشده هم می‌تواند ثابت باشد (به عنوان مثال پیشگویی برنده بعدی جایزه اسکار^{۱۶۱}، خود جواب را تغییر نمی‌دهد)، هم اینکه جواب می‌تواند شناور باشد یعنی عمل بعدی شما جواب را تغییر می‌دهد (مثل برگشت سرمایه‌گذاری شما که خود در جواب نهایی موثر است) [32].

۳-۳. خوشبندی خردمند با استفاده از آستانه‌گیری



شکل ۱-۳. چهارچوب الگوریتم خوشبندی خردمند با استفاده از آستانه‌گیری

شکل ۱-۳ چهارچوب الگوریتم خوشبندی خردمند با استفاده از آستانه‌گیری را نشان می‌دهد. در این روش کل داده به صورت مستقیم در اختیار تمام الگوریتم‌های پایه قرار می‌گیرد. روند اجرای الگوریتم بدین گونه است که ابتدا اولین الگوریتم پایه اجراشده و نتیجه آن پس از ارزیابی پراکندگی

^{۱۶۱} Oscars Awards

(باید توجه داشت چون اولین الگوریتم است هیچ الگوریتمی جهت ارزیابی درجه استقلال در بخش الگوریتم‌های انتخاب شده وجود ندارد و نتیجه ارزیابی استقلال کاملاً مستقل خواهد بود) به بخش الگوریتم‌های انتخاب شده که ما آن را با عنوان جامعه خردمند می‌شناسیم اضافه می‌شود. سپس نوبت الگوریتم بعدی است که پس از تولید نتیجه‌ی به ارزیابی درجه استقلال و میزان پراکندگی آن می‌پردازیم و در صورتی که نتایج ارزیابی از میزان آستانه تعیین شده بیشتر باشد افزار تولید شده به داخل جامعه خردمند اضافه خواهد شد. در این روش در صورت رد نتیجه به دست آمده در هر بخش فرآیند ارزیابی نتیجه الگوریتم متوقف شده و به سراغ الگوریتم پایه بعدی خواهیم رفت. در این تحقیق برخلاف روش‌های پیشین خوشبندی ترکیبی، کل افزار به دست آمده از یک الگوریتم خوشبندی پایه را در صورت داشتن شرایط لازم وارد مجمع می‌کنیم و این‌گونه اصالت جواب حفظ می‌شود. مهم‌ترین تفاوت این روش با روش‌های قبلی را می‌توان موارد زیر دانست:

اولین تفاوت این روش نحوه ارزیابی الگوریتم خوشبندی است که در این روش پس از اجرای هر الگوریتم پایه، استقلال و پراکندگی آن نسبت به سایر الگوریتم‌های داخل مجمع محاسبه می‌شود و در صورت داشتن شرایط وارد مجمع می‌شود.

دوم اینکه در اینجا به طور غیرمت مرکز الگوریتم‌ها عمل می‌کنند و الگوریتمی که بدون کیفیت، پراکندگی و استقلال باشد قادر به ورود در مجمع نیست. لذا خطاهای غیر هم جهت به وجود آمده در این روش حذف شده، و آثار جواب‌های هم جهت در مجمع بر روی هم افزوده خواهند شد که این کاملاً منطبق بر اصول حاکم بر خرد جمعی می‌باشد.

سوم اینکه چون بعد از رسیدن جمعیت مجمع به تعداد مورد نظر ما، هیچ گزینش دیگری برای تولید جواب نهایی نیاز نیست کیفیت نتایج نهایی حفظ می‌شود.

ادعاهای مطرح شده در این بخش پس از توضیح روش کار الگوریتم به صورت کامل و واضح در بخش بررسی مکانیزم بازخورد مورد مطالعه قرار می‌گیرد. در ادامه به تشریح تعاریف چهارگانه خرد جمعی مطابق با ادبیات خوشبندی ترکیبی خواهیم پرداخت.

۱-۳-۳. روش ارزیابی پراکندگی نتایج

در مورد پراکندگی آراء باید گفت چون ما در خوشبندی ترکیبی با داده‌ها و نتایج خوشبندی اولیه سر و کارداریم از واژه پراکندگی نتایج اولیه استفاده می‌کنیم و بر اساس این فرض و تعریف سورویکی از تنوع آراء آن را به صورت زیر بازنویسی می‌کنیم:

هر الگوریتم خوشبندی پایه باید به طور جداگانه و بدون واسطه به داده‌های مسئله دسترسی داشته و آن را تحلیل و خوشبندی کند حتی اگر نتایج آن غلط باشد.

در اینجا نتایج غلط موجب کشف عدم تنوع و جلوگیری از تکرار یک جواب خاص خواهد شد. ما در این تحقیق بر اساس معیار APMM (رابطه ۲۹-۲) معیاری جدید جهت سنجش پراکندگی نتیجه هر الگوریتم خوشبندی پایه ارائه می‌دهیم. در این تحقیق برای محاسبه مقدار پراکندگی یک خوشه از AAPMM (رابطه ۳۰-۲) استفاده می‌کنیم چون این معیار هم از لحاظ پیچیدگی زمانی سریع‌تر از NMI می‌باشد و هم مشکل تقارن آن را ندارد. معیاری که این تحقیق جهت سنجش پراکندگی نتیجه افزای یک خوشبندی پایه معرفی کرده است A3 نام دارد که میانگین وزن‌دار AAPMM می‌باشد که به شرح زیر است:

$$A3(p) = \frac{1}{n} \sum_{i=1}^k n_i \times AAPMM(C_i) \quad (1-3)$$

در رابطه (۳) n_i تعداد اعضای خوشه C_i و n تعداد اعضای کل خوشه‌ها و K تعداد افزاهای الگوریتم پایه می‌باشد. در این تحقیق ما مقدار آستانه dT را برای سنجش میزان پراکندگی الگوریتم خوشبندی استفاده می‌کنیم که همواره بین صفر و یک می‌باشد. پراکندگی از رابطه (۲-۳) محاسبه خواهد شد:

$$Diversity(P_i) = 1 - A3(P_i) \quad (2-3)$$

بنابراین مطابق با تعاریف بالا یکی از شرایط ورود نتیجه‌ی یک خوشبندی به مجمع رابطه (۳-۳) می‌باشد که ما آن را شرط پراکندگی می‌نامیم:

$$Diversity(C) \geq dT \quad (3-3)$$

۳-۲. روش ارزیابی استقلال الگوریتم‌ها

طبق تعریف سورویکی استقلال یعنی نتیجه رأی باید تحت تأثیر فرد یا گروه مشخصی نباشد با نگاشت این تعریف با ادبیات خوشه‌بندی ترکیبی تعریف استقلال در خوشه‌بندی را به صورت زیر بازنویسی می‌کنیم :

روش تحلیل هر یک از خوشه‌بندی‌های پایه باید تحت تأثیر روش‌های سایر خوشه‌بندی‌های پایه تعیین شود، این تأثیر می‌تواند در سطح نوع الگوریتم (گروه) یا پارامترهای اساسی یک الگوریتم خاص (افراد) باشد.

تنها واژه گنگ تعریف بالا "تعیین شدن" می‌باشد، تعیین شدن در اینجا یعنی در صورتی که یک افزار جدید تولیدشده توسط یک الگوریتم پایه بخواهد وارد مجمع (جامعه خردمند) شود باید مستقل بودن آن نسبت به سایر خوشه‌بندی‌های مجمع چک شود. برای مثال اگر ما دو الگوریتم از نوع FCM و $K-means$ داشته باشیم آنگاه چون نوع تصمیم‌گیری‌های این دو الگوریتم (روش رسیدن به نتیجه در دو الگوریتم) باهم متفاوت و مستقل است نتایج این دو خوشه‌بندی حتی در صورت برابر بودن از هم مستقل و قابل اتکا می‌باشد. در مثالی دیگر اگر دو با خوشه‌بندی پایه از نوع $K-means$ داشته باشیم و نتایج مشابه باشد و پارامترهای اساسی تصمیم‌گیری در الگوریتم برای مثال مراکز تصادفی خوشه‌ها برابر یا اختلاف ناچیزی داشته باشند آنگاه این دو خوشه‌بندی به علت استفاده از روش مشابه به همدیگر وابسته می‌باشند. بنابراین می‌گوییم در هنگام واردکردن افزارهای یک الگوریتم خوشه‌بندی پایه در مجمع، باید میزان استقلال آن از مقدار آستانه بیشتر باشد. بنا بر تعاریف بالا می‌توان درجه استقلال دو الگوریتم را با دو شرط زیر محاسبه کرد:

اول، اگر دو افزار به دست آمده، از دو الگوریتم غیر هم نام باشند به خاطر اینکه مکانیزم کار آن دو الگوریتم متفاوت است از یکدیگر مستقل هستند.

دوم، اگر دو افزار به دست آمده، از دو الگوریتم هم نام باشند درجه استقلال آن با توجه به پارامترهای اساسی آن دو الگوریتم محاسبه می‌شود.

در این تحقیق فرآیند محاسبه استقلال دو الگوریتم توسطتابع "استقلال افزایهای پایه"^{۱۶۲} محاسبه می شود که ما آن را به اختصار BPI می نامیم که در شکل ۲-۳ شبه کد آن نمایش داده شده است.

Function BPI (P1, P2) Return Result

```
If (Algorithm-Type (P1) == Algorithm-Type (P2) then  
    Result = 1 - Likeness (Basic-Parameter (P1), Basic-Parameter (P2))  
Else  
    Result = 1  
End if
```

End Function

شکل ۲-۳. محاسبه درجه استقلال دو خوشبندی

در شکل ۲-۳ پارامترهای ورودی P1 و P2 مشخصات کامل افزایهای دو الگوریتم خوشبندی ای هستند که ما قرار است میزان استقلال آنها را محاسبه کنیم و تابع Algorithm-Type نوع (اسم) الگوریتم های خوشبندی را بر می گرداند برای مثال $K-means$ و یا FCM و غیره. ماتریس پارامترهای اساسی یا Basic-Parameter شامل پارامترهای مهم هر الگوریتم می باشد که تحت تأثیر آن الگوریتم های خوشبندی شروع به حل مسئله می کنند برای مثال نقاط اولیه مراکز، مقادیر تصادفی و غیره. این مقادیر می توانند بر اساس دو عامل تعریف شوند: اول طبیعت مسئله و فضای حل آن و دوم نوع الگوریتم و مکانیزم حل آن. به عنوان یک مثال اضافه می توان به مراکز تصادفی اولیه در الگوریتم $K-means$ اشاره کرد که هرچه این مقادیر در ابتدا دارای فاصله بیشتر باشند و جواب های نهایی حل شده توسط این الگوریتم به هم شبیه تر باشند می توان احتمال بالاتری برای درستی این جواب ها در نظر گرفت. این تحقیق برای مقایسه ماتریس های پارامترهای اساسی با یکدیگر از تابع مکاشفه ای Likeness استفاده می کند که در ادامه آن را شرح می دهیم.

در محاسبه تابع Likeness ما فرض می کنیم که MaxDis بیشترین مقدار در ماتریس های فاصله می باشد (در این تحقیق ما از معیار فاصله اقلیدسی برای محاسبه فاصله استفاده کردیم ولی در روش پیشنهادی این تحقیق می توان از هر معیار فاصله دیگری نیز استفاده کرد) و ماتریس های MAT_A و MAT_B ماتریس های پارامترهای اساسی دو الگوریتم خوشبندی که قرار است درجه استقلال آنها

¹⁶² Basic-Partition-Independence

نسبت به هم محاسبه شود می‌باشد (بخشی از اطلاعات ورودی‌های P1 و P2). در این صورت ماتریس شباهت $LMAT_t$ برای MAT_A و MAT_B فرض خواهد شد که در آن Sim_t مقدار کمینه این ماتریس می‌باشد. با حذف سطر و ستونی که در آن مقدار Sim_t وجود دارد ماتریس $LMAT_{t+1}$ ایجاد شده که از طریق مشابه می‌توان Sim_{t+1} را محاسبه کرد. این کار آن قدر تکرار خواهد شد تا کل داده‌های ماتریس‌های $LMAT$ حذف شوند. مقدار تابع Likeness بر اساس تعریف بالا از رابطه ۴-۳ محاسبه می‌شود:

$$Likeness = 1 - \left(\frac{1}{MaxDis} \sum_{t=0}^n Sim_t \right) \quad (4-3)$$

لازم به ذکر است که در تعاریف بالا $LMAT_0$ یک ماتریس $n \times n$ است که n تعداد پارامترهای اساسی در الگوریتم می‌باشد. استقلال هر افزار در جامعه خردمند با رابطه ۵-۳ محاسبه می‌شود:

$$Independence(P) = \frac{1}{M} \sum_{i=1}^M BPI(P, P_i) \quad (5-3)$$

که در آن M تعداد اعضای جامعه خردمند و تابع BPI با استفاده از شبکه کد شکل ۲-۳ می‌باشد. بنا بر تعاریف بالا شرط ورود یک افزار به جامعه خردمند بزرگی درجه استقلال آن از مقدار آستانه iT می‌باشد. رابطه ۶-۳ این شرط را نشان می‌دهد که ما آن را با عنوان شرط استقلال می‌شناسیم. لازم به ذکر است که مقدار آستانه $1 \leq iT \leq 0$ می‌باشد.

$$Independence(C) \geq iT \quad (6-3)$$

به عنوان آخرین نکته باید اشاره کرد که استقلال یک معیار پراکندگی نیست به خاطر اینکه معیارهای پراکندگی برای ارزیابی نتایج به دست آمده از خوشبندی‌های اولیه مورد استفاده قرار می‌گیرند ولی معیار استقلال فرآیند تولید نتایج را کنترل می‌کند، این عمل با مدیریت پارامترهای اساسی در هر الگوریتم محقق می‌شود. علاوه بر آن، استقلال می‌تواند احتمال درستی الگوهای مشابه را محاسبه کند ولی از دیدگاه معیارهای پراکندگی دو الگوی مشابه دارای پراکندگی بسیار پایین می‌باشند. از طرف دیگر با این که معیار استقلال می‌تواند احتمال درستی جواب را بررسی کند ولی نمی‌تواند تضمین کند تا جواب نهایی به دست آمده از پراکندگی لازم برخوردار باشند و فقط می‌تواند پراکندگی آن را تا حدی بهبود ببخشد از این روی در این تحقیق معیار استقلال به عنوان معیاری مکمل برای معیار پراکندگی معرفی شده است نه به جای آن.

۳-۳-۳. عدم تمرکز در بخش‌های سازنده خوشبندی ترکیبی

سورویکی موارد لازم برای ایجاد یک مجمع خردمند را این‌گونه بیان می‌کند. "شرایط لازم برای جامعه خردمند شامل پراکندگی، استقلال و نوعی خاصی از عدم تمرکز می‌باشد." با توجه به توضیحات سورویکی در مورد عدم تمرکز و مثل‌هایی که از سازمان^{۱۶۳} CIA یا سیستم عامل لینوکس^{۱۶۴} و مشکل طراحی شاتل^{۱۶۵} در ناسا^{۱۶۶} ذکر می‌کند در مورد عدم تمرکز باید گفت این معیار، یک معیار کیفی است. روش کنترل این متريک باید در سطح بخش‌های سازنده خوشبندی خرد جمعی باشد [55]. عدم تمرکز را بر اساس تعاريف بالا در خوشبندی ترکیبی اين‌گونه تعريف می‌کنيم:

ارتباط بين بخش‌های مختلف خوشبندی خرد جمعی باید به گونه‌ای باشد تا بر روی عملکرد خوشبندی پايه تأثيری ایجاد نکند تا از اين طریق هر خوشبندی پايه شناس این را داشته باشد تا با شخصی سازی و بر اساس دانش محلی خود بهترین نتیجه ممکن را آشکار سازد.

در اينجا بهترین نتیجه ممکن يعني نتیجه‌اي که داراي ميزان استقلال و پراکندگی بهينه باشد و در مجمع پذيرفته شود. نتیجه مهمی که از تعاريف بالا می‌توان دریافت اين است که عدم تمرکز در خرد جمعی با روش ارتباط بخش‌های مختلف باهم در خوشبندی خرد جمعی در رابطه مستقيم می‌باشد از اين رو ما در طراحی سیستم خوشبندی ترکیبی باید شرایط زير را جهت حفظ عدم تمرکز رعایت کنيم:

۱- تعداد الگوريتم‌های پايه شركت‌کننده باید بيشتر از يك الگوريتم باشد.

۲- روش ورود يك الگوريتم پايه به مجمع باید طوري باشد تا نتایج نهايی تحت تأثير خطاهای آن قرار نگیرد يا به عبارتی نباید روش تصمیم‌گیری در مورد جواب نهايی متمرکز باشد.

¹⁶³ Central Intelligence Agency

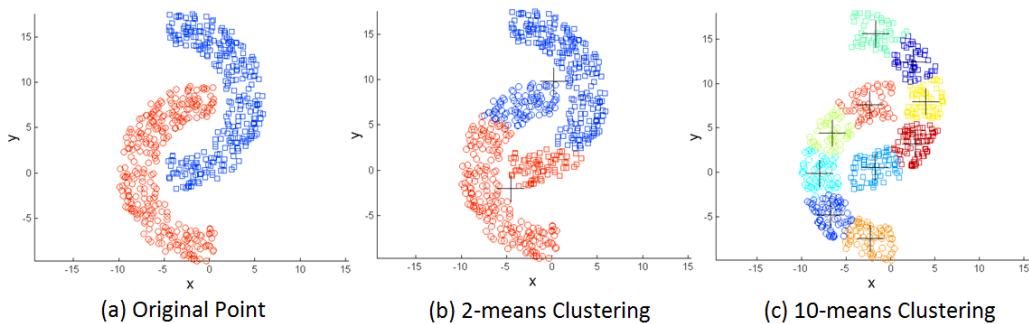
¹⁶⁴ Linux

¹⁶⁵ Shuttle

¹⁶⁶ National Aeronautics and Space Administration

۳- مقدار آستانه cT ضریب عدم تمرکز نامیده می‌شود که آن به عنوان ضریب برای تعداد خوشه‌های الگوریتم پایه استفاده می‌شود. در این روش تعداد خوشه‌ها در خوشبندی پایه از $cT \times k$ متغیر می‌باشد.

مطابق تعاریف بالا، ضریب cT که عضو اعداد طبیعی است، وقتی داده دارای پیچیدگی‌های خاصی است و الگوریتم خوشبندی پایه نمی‌تواند الگوهای آن را شناسایی کند، می‌تواند دقیقت در جواب نهایی را بپسورد بخشد. این معیار می‌تواند با افزایش تعداد خوشه در الگوریتم خوشبندی پایه پیچیدگی آن را کاهش دهد [68] و الگوهای پیچیده داده را به الگوهای ساده کوچکتر تبدیل کند که راحت‌تر توسط الگوریتم قابل تشخیص باشد (به ویژه برای الگوریتم‌های مبتنی بر مرکز خوشه^{۶۷}). این روش به جای پیدا کردن یک راه حل مجتمع برای هر مسئله پیچیده سعی می‌کند تا این مسئله را به مسائل کوچک‌تر و با الگوهای ساده‌تر تبدیل کند و آنها را حل نماید. برای مثال حل داده‌هایی با شکل غیر کروی توسط الگوریتم‌های مبتنی بر مرکز خوشه همانند الگوریتم پایه K -means از کاربردهای این متریک می‌باشد. شکل ۳-۳ نشان‌دهنده تأثیر عدم تمرکز بر روی داده Halfring می‌باشد:



شکل ۳-۳. تأثیر عدم تمرکز بر روی پیچیدگی داده

بخش (a) شکل ۳-۳ نشان‌دهنده شکل داده Halfring می‌باشد که در آن رنگ آبی و قرمز دو کلاس این داده است. بخش (b) شکل ۳-۳ نتیجه خوشبندی این داده را با استفاده از K -means به ازای $k = 2$ (مقدار k برابر با تعداد واقعی کلاس داده می‌باشد) نشان می‌دهد. همان‌طور که بخش (b) نشان می‌دهد الگوریتم K -means قادر به حل این داده نمی‌باشد. بر اساس روش پیشنهادی این تحقیق اگر $cT = 5$ فرض شود آنگاه این داده پیچیده به بخش‌های ساده کوچک‌تر تبدیل می‌شود که

^{۶۷} Ccenter-Based-Clustering algorithms

به راحتی با استفاده از الگوریتم $K-means$ قابل تشخیص خواهد بود. بخش (c) شکل ۳-۳ نشان‌دهنده خروجی به ازای تعداد خوش $cT \times k = 10$ می‌باشد.

در انتها باید به این نکته اشاره کرد که پیاده‌سازی درست عدم مرکز نیاز به رعایت نکاتی در تمامی بخش‌های تشکیل‌دهنده خوش‌بندی ترکیبی دارد که ما در ادامه در بخش "بررسی تأثیر مکانیزم بازخورد"^{۱۶۸} در کیفیت نتیجه نهایی "آن را کاملاً بررسی می‌کنیم.

۴-۳-۳. مکانیزم ترکیب مناسب

مطابق با تعاریف خرد جمعی روشی مناسب جهت ادغام نتایج به شکل زیر تعریف می‌کنیم:

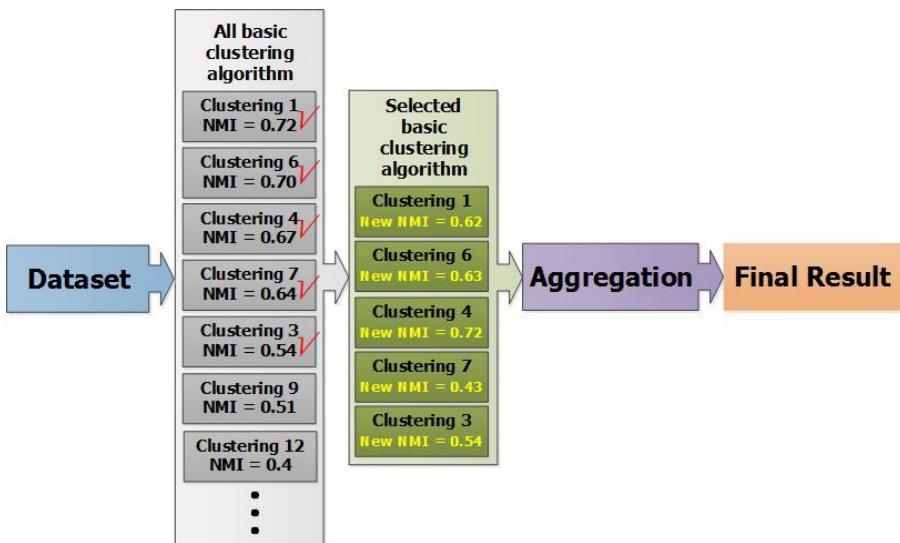
باید مکانیزمی وجود داشته باشد که بتوان توسط آن نتایج اولیه الگوریتم‌های پایه را با یکدیگر ترکیب کرده و به یک نتیجه نهایی (نظر جمعی) رسید.

در این تحقیق ما از روش ماتریس همبستگی که پیش‌تر به آن اشاره شد برای تشکیل نتیجه نهایی با استفاده از نتایج اولیه بهره می‌بریم. برای این منظور از رابطه ۴۲-۲ استفاده خواهیم کرد. لازم به ذکر است که نتایج تجربی چندین مقاله کارایی این روش را اثبات می‌کند [1, 2, 8, 9, 19, 23].

۴-۳-۴. بررسی تأثیر مکانیزم بازخورد در کیفیت نتیجه نهایی

در روش‌های پیشین خوش‌بندی ترکیبی ابتدا تمامی نتایج خوش‌بندی اولیه تولید شده و پس از آن کلیه این نتایج بر اساس یک معیار ارزیابی خوش‌بندی همانند NMI ارزیابی می‌شود و بر اساس این نتایج ارزیابی افزارهای (در برخی موارد خوش‌ها) بهتر انتخاب می‌شود. شکل ۴-۳ نشان‌دهنده این فرآیند است. همان‌گونه که در این شکل مشاهده می‌شود اگر چه افزارهای به دست آمده در مجموعه کل افزارها دارای بیشترین مقدار متریک مورد ارزیابی می‌باشد ولی پس از انتخاب مقدار این ارزیابی‌ها تغییر کرده و در بیشتر موارد کاهش می‌یابند. این بدان معنی است که در روش‌های پیشین اگر چه با کیفیت‌ترین (در بیشتر موارد پراکنده‌ترین) نتایج انتخاب می‌شوند ولی این مکانیزم نمی‌تواند دوام این کیفیت را تضمین کند.

¹⁶⁸ Feedback Mechanism



شکل ۳-۳. تأثیر انتخاب افزارها در خوشبندی ترکیبی مبتنی بر انتخاب بر مقدار NMI ارزیابی شده.

از طرف دیگر، روش پیشنهادی این تحقیق با استفاده از مکانیزم بازخورد تعداد اعضای جامعه خردمند را به تدریج افزایش می‌دهد. در این روش پس از تولید یک نتیجه (افراز) با استفاده از الگوریتم‌های خوشبندی پایه آن را با استفاده از متريک استقلال و پراکندگی ارزیابی می‌کنیم. اگر نتایج این ارزیابی قابل قبول باشد کل افراز به دست آمده به مجموعه جواب‌های انتخاب شده یا همان جامعه خردمند اضافه می‌شود در غیر این صورت به طور خودکار حذف می‌شود. این فرآيند برای تمامی نتایج تکرار خواهد شد. قابل به ذکر است که اين روند موجب می‌شود تا كيفيت نتایج اوليه به دست آمده برای توليد نتیجه نهايی حفظ شود زيرا نتایج ارزیابی ها پس از تغيير در تعداد اعضای جامعه خردمند به روز شده و بعد از اتمام فرآيند توليد نتایج هیچ گزینش دیگری صورت نمی‌گيرد.

۳-۶. شبکه خوشبندی خردمند با استفاده از آستانه‌گيری

همان طور که پيش‌تر اشاره شد شکل ۱-۳ چهارچوب روش پیشنهادی اول این تحقیق را ارائه می‌دهد. این فرآيند با توليد نتایج اوليه شروع شده و با ارزیابی پراکندگی و استقلال نتایج اقدام به تولید جامعه خردمند می‌کند. در نهايىت با استفاده از روش انباشت مدارك افرازهای جمع‌آوری شده در جامعه خردمند با يكديگر ادغام شده و نتیجه نهايى را توليد می‌کند. شکل ۴-۳ نشان‌دهنده شبکه کد روش پیشنهادی اول است. در اين شکل Kb تعداد خوشبندی‌ها در الگوریتم پایه می‌باشد و تابع Generate-Basic-Algorithm نتایج اوليه (افرازهای) را با استفاده از الگوریتم‌های خوشبندی‌های پایه را توليد می‌کند. دو تابع Independence و Diversity به ترتیب برای ارزیابی پراکندگی و استقلال به

کار می‌رود. تابع Make-Correlation-Matrix ماتریس همبستگی را برای تولید نتیجه نهایی با استفاده از نتایج اولیه بر اساس رابطه $56-2$ تولید می‌کند. برای تولید دندوگرام از ماتریس همبستگی ما از الگوریتم پیوندی میانگین استفاده کردایم چون نتایج تجربی این تحقیق که در بخش ارزیابی ارائه می‌شود نشان داده است که این روش بهترین دقیقت را دارد. در اینجا تابع Average-Linkage نشان‌دهنده الگوریتم پیوندی میانگین است و همچنین تابع Cluster بر اساس تعداد خوشه تعیین شده نتیجه نهایی را از روی دندوگرام تشکیل می‌دهد.

Function WOCCE (Dataset, Kb, iT, dT, cT) Return [Result, nCrowd]

Initialized nCrowd to zero

While we have base cluster

[idx, Basic-Parameter] = Generate-Basic-Algorithm (Dataset, Kb*cT)

If (Independence (Basic-Parameter) > iT) **then**

If (Diversity (idx) > dT) **then**

Insert idx to Crowd-Partitions

Crowd = Crowd + 1

End if

End if

End while

Co-Acc = Make-Correlation-Matrix (Crowd-Partition)

Z = Average-Linkage (Co-Acc)

Result = Cluster (Z, Kb)

شکل ۳-۴. شبکه کد خوشه‌بندی خردمند با استفاده از آستانه‌گیری

جدول ۳-۱ نشان‌دهنده نگاشت لغات لاتین در ادبیات حوزه خوشه‌بندی ترکیبی به نظریه خرد جمعی می‌باشد.

جدول ۳-۱. نگاشت لغات لاتین در خوشبندی ترکیبی به نظریه خرد جمعی

WOC Terminology	Cluster Ensemble Terminology
Primary opinion	Primary partition
People	Base algorithm
Wise crowd	Primary clustering results
Diversity of Opinion	Diversity of primary clustering results
Opinion independence	Independence of clustering algorithms that generate primary partitions
Decentralization	Decentralization in cluster generation

۴-۳. خوشبندی خردمند مبتنی بر گراف استقلال الگوریتم

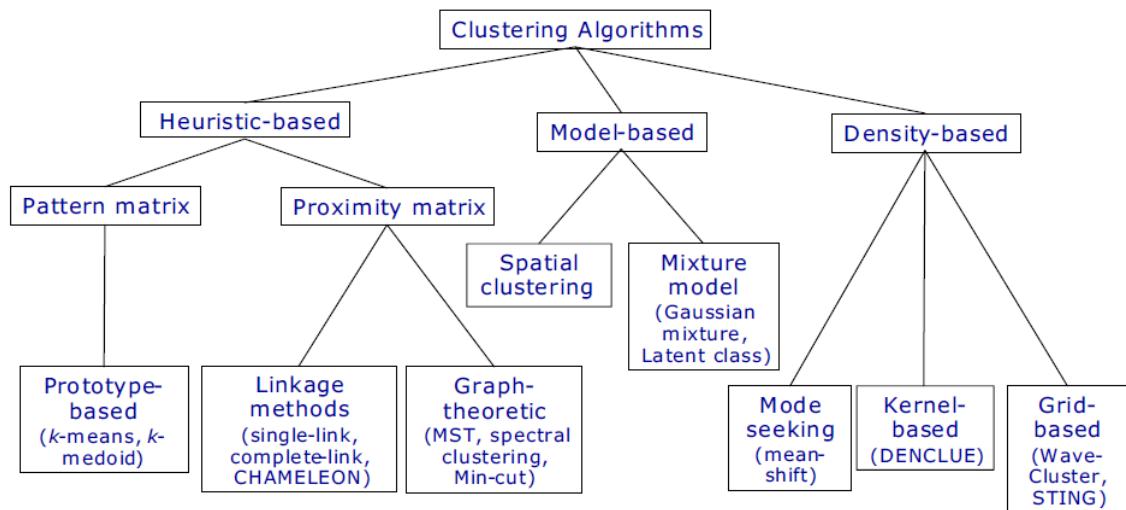
روش پیشنهادی دوم این تحقیق سعی دارد تا دو بخش از روش اول را بهبود بخشد. این روش در ابتدا این ایده را بررسی می‌کند که الگوریتم‌های غیر هم نام کاملاً مستقل نیستند. در این راستای برای محاسبه درجه استقلال دو الگوریتم با استفاده از ایده تبدیل کد به گراف در تست نرم‌افزار به ارائه روشی با عنوان مدل‌سازی گراف استقلال الگوریتم می‌پردازیم. با مقایسه گراف‌های به دست آمده (درجه استقلال) در این روش می‌توان یک وزن برای احتمال صحت جواب‌های به دست آمده پیشنهاد داد. از این روی در این روش به جای ارزیابی و آستانه‌گیری از استقلال الگوریتم‌ها، آن‌ها را به عنوان وزنی برای ترکیب نتایج در نظر می‌گیریم بدین وسیله نیاز به آستانه‌گیری برای استقلال الگوریتم‌ها از بین می‌رود. جهت ترکیب نتایج اولیه در این روش رابطه‌ای جدید بر اساس رابطه ۵۶-۲ با عنوان روش انباشت مدارک وزن‌دار یا به اختصار WEAC^{۱۶۹} معرفی می‌شود.

۴-۴-۱. بررسی مکانیزم حل مسائل توسط الگوریتم‌های خوشبندی

در سال‌های اخیر دسته‌بندی‌های زیادی برای الگوریتم‌های خوشبندی در کتب و مقالات مختلف ارائه شده است [37, 71, 72, 68]. یکی از دیدگاه‌های غالب در این بحث روشی است که توسط جین و همکاران [37] پیشنهاد شده است. در این روش، دسته‌بندی الگوریتم‌ها بر اساس تابع

¹⁶⁹ Weighted Evidence Accumulation Clustering

هدف^{۱۷۰} الگوریتم‌ها صورت می‌گیرد. شکل ۵-۳ دسته‌بندی پیشنهادشده توسط جین و همکاران را نشان می‌دهد.



شکل ۵-۳. دسته‌بندی الگوریتم‌های خوشبندی [37].

نتایج تجربی جین و همکاران نشان می‌دهد که روش عملکرد هر دسته از الگوریتم‌های خوشبندی که در یک گروه قرار دارند (دارای تابع هدف مشابه هستند) بر روی یک سری داده‌های خاص تقریباً یکسان است [37]. از طرف دیگر، می‌توان بسیاری از الگوریتم‌های خوشبندی را یافت که روشی توسعه‌یافته از روی یک الگوریتم پایه می‌باشد. شاید معروف‌ترین مثال‌هایی که بتوان از این‌گونه الگوریتم‌ها نام برد الگوریتم‌های سری *K-means* می‌باشند که جین [74] پنجاه سال تکامل و انواع آن را بررسی کرده است و یا سری الگوریتم‌های پیوندی که ما پیش‌تر به چند نمونه از آن اشاره کرده‌ایم. در بیشتر این سری از الگوریتم‌ها فرآیند کلی حل مسئله ثابت است و تنها بخشی از الگوریتم اضافه، تغییر، حذف و یا بهبود داده شده است. از این روی با توجه به این رابطه خاص بین برخی از انواع الگوریتم‌ها روش پیشنهادی دوم این تحقیق فرآیندی را برای مدل‌سازی شیوه کار الگوریتم‌های خوشبندی پیشنهاد می‌دهد تا بر اساس آن استقلال و یا وابستگی الگوریتم‌های خوشبندی به نحوی دقیق‌تر از روش پیشنهادی اول، جهت کنترل تأثیرات آن‌ها بر روی تولید نتایج اولیه خوشبندی ترکیبی مورد توجه قرار گیرند.

¹⁷⁰ Objective Function

برای ارزیابی و سنجش درجه استقلال و یا وابستگی الگوریتم‌های خوشبندی ابتدا نیاز است آنها را مدل‌سازی کنیم. این مدل‌سازی باید بر شیوه‌ای استوار باشد که بتوان فقط و فقط عواملی که بر استقلال و یا وابستگی یک الگوریتم تأثیر دارند همانند شیوه کار الگوریتم، مولدهای اعداد تصادفی، روابط ریاضی، توابع مکاشفه‌ای و غیره را بتوان بدون کمترین تأثیری از سایر بخش‌های غیر مرتبط همانند کدهای ورودی، خروجی و نمایش و غیره مدل‌سازی کند. در این تحقیق بر اساس ایده روش مدل‌سازی کد برنامه به گراف که در تست نرم‌افزار^{۱۷۱} استفاده می‌شود روشی جهت مدل‌سازی فرآیند کار الگوریتم برای ارزیابی استقلال پیشنهاد می‌شود که ما آن را با عنوان "مدل‌سازی گراف استقلال الگوریتم" می‌شناسیم. در ادامه به بررسی این روش خواهیم پرداخت.

۲-۴-۳. مدل‌سازی گراف استقلال الگوریتم

تست نرم افزا یکی از مهم‌ترین بخش‌های ساخت نرم‌افزار می‌باشد که تقریباً صفت درصد از کل هزینه تولید یک نرم‌افزار به آن تخصیص داده می‌شود. یکی از روش‌های تست نرم افزا مدل‌سازی نرم‌افزار می‌باشد که به چهار بخش مدل‌های مبتنی بر ^{۱۷۲} نحو کد برنامه، فضای ورودی، منطق عملکرد و گراف تقسیم می‌شود. از بین این روش‌ها، مدل‌سازی مبتنی بر گراف می‌تواند یک دیدگاه نمایش گرافیکی از روی کد منبع، طراحی، مشخصات یا موردهای استفاده^{۱۷۳} ارائه دهد از این روی این روش برای بررسی مکانیزم کار یک الگوریتم بسیار مفید می‌باشد [75]. در این تحقیق ما از مفاهیم و ایده مدل‌سازی مبتنی بر گراف در تست نرم‌افزار جهت محاسبه درجه استقلال الگوریتم‌های خوشبندی استفاده می‌کنیم. قبل از اینکه به تشریح روش ساخت گراف استقلال بپردازیم باید به دو سؤال در ساخت این گراف کمی دقیق‌تر پاسخ داد.

اول، گراف استقلال باید بر اساس چه منبعی ساخته شود؟

برای پاسخ به این سؤال باید یادآور شد که گراف استقلال فرآیند روش حل مسئله در الگوریتم خوشبندی را مدل‌سازی می‌کند لیکن باید آن را از روی کد و یا شبه کد پیاده‌سازی الگوریتم ساخت تا بتوان رابطه‌ی بین این کدها را شناسایی کرد و از روی آنها استقلال الگوریتم را ارزیابی کنیم.

¹⁷¹ Software Test

¹⁷² Syntax

¹⁷³ Use Case

دوم، آیا کل کدهای پیاده‌سازی یک الگوریتم خوشه‌بندی باید در مدل‌سازی بکار رود؟

همان طور که پیش‌تر نیز اشاره شد هر الگوریتم شامل بخش‌هایی برای تهییه و ورودی، خروجی و نمایش داده به کاربر می‌باشد. علاوه بر آن همیشه کدهای الگوریتم شامل بخش‌هایی برای تعریف متغیرها، ثابت‌ها و غیره هستند که در بیشتر مواقع فایده‌ای برای ارزیابی درجه استقلال ندارند. این موارد در شبه کدها کمتر به چشم می‌خورند ولی با این حال همیشه وجود دارند. از آنجایی که روش پیشنهادی این تحقیق به این تعاریف حساس می‌باشد. از این روی در این تحقیق روشی برای هرس کردن کد و تبدیل آن به قالبی مناسب تشخیص استقلال با عنوان "زبان استقلال الگوریتم خوشه‌بندی"^{۱۷۴} یا به اختصار CAIL پیشنهاد خواهیم کرد.

۳-۴-۲-۱. زبان استقلال الگوریتم خوشه‌بندی

در زبان استقلال الگوریتم خوشه‌بندی به جای استفاده از کد یا شبه کد الگوریتم‌ها از نمادهای توافقی استفاده می‌شود. مهمترین دلایل این کار را می‌توان موارد ذیل دانست. اولاً، از آن جایی که معمولاً کدها و یا شبه کدها به زبان استانداردی نوشته نمی‌شوند لذا برای مقایسه باید تمامی آن‌ها را به یک شکل همگن کرد. علاوه بر آن چون در بسیاری از موارد معادلات ریاضی و شبه کدها در مقالات واضح بیان نمی‌شوند اگر قرار باشد بر اساس آن‌ها به مدل‌سازی یک الگوریتم بپردازیم باید به شیوه‌ای عمل کنیم تا حساسیت به جزئیات پیاده‌سازی کم شود. در این تحقیق پس از طی یک سری فرآیند استاندارد تمامی کدها به کدی استاندارد برای ارزیابی استقلال تبدیل می‌شود. این فرآیند به شرح زیر است:

۱- ابتدا تمامی کدهای اضافی از جمله تعریف متغیرها و ثابت‌ها، توضیحات اضافی، دستورات مربوط به ورودی، خروجی و نمایش اطلاعات، تمامی دستورات مربوط به کترل حلقه‌ها و شروط در صورتی که تغییری در شرایط اجرای الگوریتم ایجاد نکنند و سایر کدهایی که در فرآیند خوشه‌بندی داده نقشی ندارند را حذف می‌کنیم. برای مثال در صورتی که پیاده‌سازی تابعی خاص که در کد اصلی از آن استفاده شده است همانند معیارهای

^{۱۷۴} Clustering Algorithm Independence Language

ارزیابی همچون NMI و APMM یا غیره در کد اصلی وجود داشته باشد پیاده‌سازی آن را حذف می‌کنیم چون در نماد گزاری به راحتی می‌توان کل آن را با یک نماد توافقی نشان داد.

۲- چون بخش منطقی عملگرهای شرطی و حلقه‌ها تأثیری در شکل گراف نمی‌گذارند آن‌ها را حذف می‌کنیم.

۳- با حذف بخش منطقی عملگرهای شرطی و حلقه‌ها استفاده از انواع عملگر شرطی (همانند If، case و elsif و غیره) و همچنین به کارگیری انواع عملگر حلقه (همانند حلقه‌های for و while و غیره) نیاز نیست. قابل به ذکر است که کلیه این دستورات برای تسريع سرعت پیاده‌سازی کد الگوریتم توسط برنامه‌نویس به کار می‌روند ولی در نهایت همگی آن‌ها جزئی از دو گروه عملگرهای شرطی و حلقه‌ها می‌باشند. چون در مدل‌سازی الگوریتم فرآیند کار الگوریتم مدل می‌شود و نه نحوه پیاده‌سازی تک‌تک کدها، از این روی دستور بکار رفته در کد اهمیتی ندارند بلکه فقط و فقط ماهیت آن‌ها مهم می‌باشد. در نتیجه این تحقیق پیشنهاد می‌کند برای سادگی مدل تنها از یک دستور قراردادی به عنوان عملگر شرط و یک دستور قراردادی برای عملگر حلقه استفاده کنیم. این تحقیق برای هر نوع شرطی از نمادهای If Else و برای هر نوع حلقه‌ای از نمادهای While Break End استفاده می‌کند.

۴- به منظور سادگی و همگن کردن کدها و همچنین تسريع در عمل مقایسه و ارزیابی آن‌ها در این تحقیق از جدولی قراردادی برای تبدیل کدهای اصلی به کدهای استاندارد استفاده می‌شود. این جدول که ما آن را با عنوان "جدول نگاشت استاندارد کد^{۱۷۵}" و یا SCMT می‌شناسیم می‌تواند با حفظ قالب اصلی خود مناسب با هر پیاده‌سازی تهیه شود. در ادامه چند قانون قراردادی را برای تولید این جدول بیان می‌کنیم.

۵- جهت وضوح بیشتر کد تهیه‌شده ابتدای آن با کلمه Begin و انتهای آن با کلمه End معین می‌شود.

به منظور سهولت در انتخاب نماد در جدول نگاشت استاندارد کد، پیشنهاد می‌شود تا دستورات ابتدای گروه‌بندی شده و برای هر دستور در گروه منحصر به فرد آن یک برچسب عددی در نظر گرفته شود. در اینجا گروه‌ها با حروف انگلیسی و برچسب‌ها در مقابل آن داخل پرانتز نوشته خواهد شد. این

¹⁷⁵ Standard Code Mapping Table

گروه‌بندی و برچسب‌گذاری می‌تواند به هر صورت دلخواهی انجام شود ولی حتماً باید برای حل یک مسئله خاص برچسب‌گذاری و گروه‌بندی تمامی الگوریتم‌ها یکتا باشند. برای مثال می‌توان گروه مولد اعداد تصادفی را با R ، گروه روابط ریاضی را با M ، روابط مکاشفه‌ای را با H و گروه تخصیص برچسب خوش‌به داده را با G نشان داد. در جدول ۲-۳ مثالی از یک نمونه بسیار ساده از جدول نگاشت استاندارد کد به تصویر کشیده شده است.

جدول ۲-۳. یک نمونه از جدول نگاشت استاندارد کد

ناماد	کد برنامه
$R(0)$	$Y = \text{RandomNumber}()$
$R(1)$	$Y = \text{RandomNumbers}(k)$
$M(1)$	$Y = \text{Sin}(X)$
$M(2)$	$Z = \text{EuclideanDistance}(X, Y)$
$M(3)$	$Z = \text{ManhattanDistance}(X, Y)$
$H(1)$	$Z = \text{NMI}(A, B)$
$H(2)$	$Z = A3(A, B)$
$G(1)$	Assign data point to cluster with nearest center
$G(2)$	Connect data point to nested with max center

بر اساس برخی از نمادهای جدول ۲-۳ و شبه کد شکل ۱۰-۲ برای الگوریتم K -means کدی با قالب استاندارد زبان استقلال الگوریتم خوش‌به دی در شکل ۶-۳ به تصویر کشیده شده است.

Begin

$R(2)$

While

$G(1)$

$M(2)$

End

End

شکل ۶-۳. کد الگوریتم K -means به زبان استقلال الگوریتم خوش‌به دی

همان طور که مشاهده می‌کنید کد شکل ۶-۳ فاقد هر گونه جزئیات اضافی و غیر مرتبط می‌باشد و به خوبی فرآیند اجرای الگوریتم به همراه دستورات موثر در استقلال آن را به بیان می‌کند. علاوه بر آن، با توجه به جدول ۲-۳ می‌توان گفت که یکی دیگر از ویژگی‌های مهم جدول نگاشت استاندارد کد

این است که در تهیه این جدول می‌توان همزمان از بخشی از کد برنامه یک الگوریتم و شبه کد الگوریتمی دیگر در کنار یکدیگر استفاده کرد بدون آن که در همگن بودن کد نهایی مشکلی پیش آید. در بخش "مدل‌سازی الگوریتم‌ها به زبان استقلال الگوریتم" جدول نگاشت استاندارد کد استفاده شده جهت مدل‌سازی الگوریتم‌های این تحقیق و کدهای مدل‌سازی شده آن الگوریتم‌ها به قالب استاندارد زبان استقلال الگوریتم خوشبندی ارائه می‌شود.

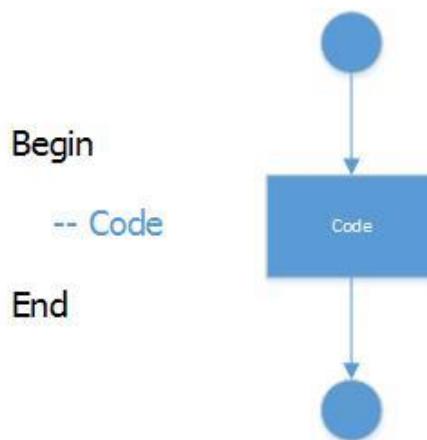
۳-۴-۲. تبدیل کد به گراف استقلال الگوریتم

جهت ساخت گراف استقلال الگوریتم از روی کدی با قالب استاندارد زبان استقلال الگوریتم از روش تبدیل کد به گراف در مباحث تست نرم‌افزار [75] استفاده می‌کنیم. چون کاربرد گراف استقلال خاص منظوره است و کد تهیه شده نیز دارای یک قالب استاندارد می‌باشد از این روی روش ساخت گراف مطرح شده در این تحقیق کاملاً مشابه روش‌های ساخت گراف در تست نرم‌افزار نیست بلکه روشی سفارشی شده بر اساس ایده تبدیل کد به گراف متناسب با نیازهای ارزیابی استقلال الگوریتم می‌باشد. در این روش، اتصالات^{۱۷۶} کد (که مطابق با تعاریف بخش قبل شامل نقاط شروع، پایان، شرط و حلقه و بخش‌های زیرمجموعه‌ی آنها می‌باشند) را طوری در نظرخواهیم گرفت که هر کدام مطابق با فرآیند کارشان به چند نود و یال‌های میان آنها تقسیم شوند. در این روش برای نشان دادن روند اجرای برنامه از گراف جهت‌دار همانند روش تست نرم‌افزار استفاده خواهیم کرد. از طرف دیگر، در روش تست نرم‌افزار کد میان اتصالات در بخش گره‌ها نشان داده می‌شوند و روی یال‌ها غالباً بخش منطقی عملگرهای شرطی و یا حلقه نوشته می‌شود ولی به خاطر این که اولاً ما بخش منطقی عملگرها را حذف کرده‌ایم ثانیاً در این ارزیابی روند اجرای کد مهم است و ثالثاً کدها به شکل هرس شده و با نمادهای استاندارد پیاده‌سازی شده‌اند، پس برای سادگی ووضوح بیشتر کد به جای نوشتن کدها در داخل گره‌ها آنها را روی یال‌های ما قبل گره که دقیقاً جریان همان کد را نشان می‌دهد نوشته و به عنوان وزن (غیر عددی) آن یال در نظرخواهیم گرفت. روش نوشتن هر بخش از این کدها به عنوان وزن در گراف باید با همان ترتیبی که در کد اصلی نوشته شده است باشد. بدیهی است که به سادگی می‌توان شکل و نمادهای گراف‌های تهیه شده برای دو الگوریتم را باهم مقایسه کرد و میزان شباهت و یا تفاوت این گراف‌ها نشان‌دهنده روش عملکرد آنها (روش حل مسئله)

¹⁷⁶ Conjunction

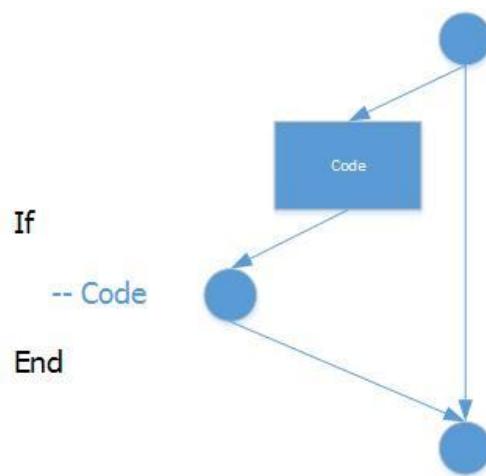
می باشد. با مقایسه نمودن این گراف‌ها می‌توان میزان استقلال الگوریتم‌های خوشبندی را ارزیابی کرد. در ادامه ابتدا روش تبدیل هر اتصال در کد را به گراف و حالت‌های خاص به وجود آمده را بررسی می‌کنیم و سپس چند مثال پرکاربرد را برای تبدیل کد به گراف نشان می‌دهیم.

گراف بخش شروع و پایان شکل ۷-۳ روش تبدیل کدهای شروع و پایان به گراف را نشان می‌دهد. در این شکل دایره‌ها، گره‌های گراف می‌باشند و شکل مربع شامل تمامی گره‌ها و یال‌هایی است که از روی کد نوشته شده بین دو کلمه کلیدی به گراف تبدیل شده است.



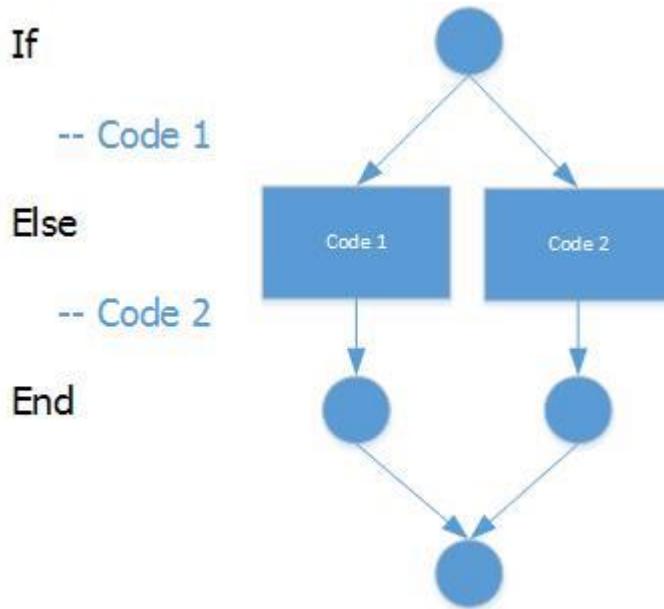
شکل ۷-۳. تبدیل کدهای شروع و پایان به گراف

گراف عملگر شرط شکل ۸-۳ نشان‌دهنده تبدیل ساده‌ترین نوع عملگر شرطی به گراف می‌باشد.



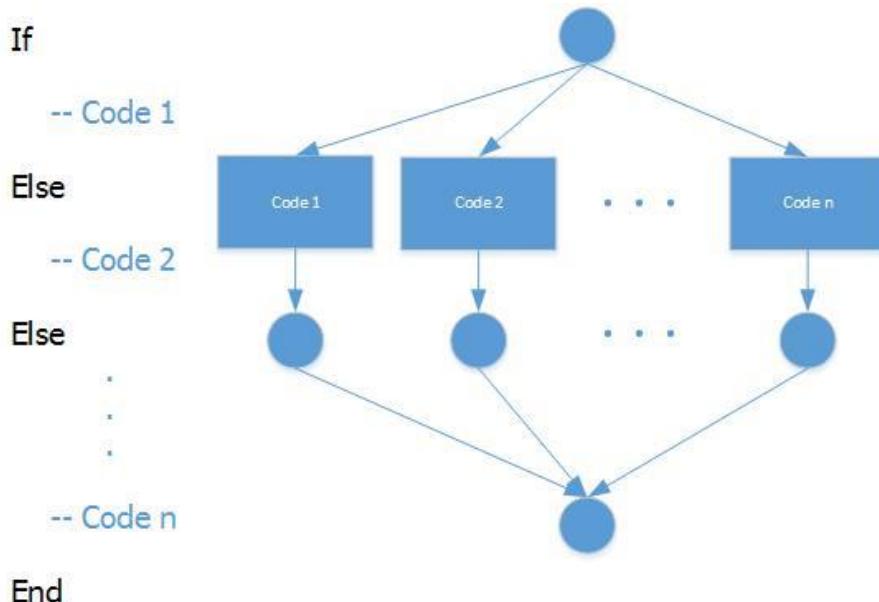
شکل ۸-۳. تبدیل عملگر شرط ساده به گراف

شکل ۹-۳ نمایش گراف معروف‌ترین نوع عملگر شرط می‌باشد که ما آن را با عنوان عملگر شرط کامل می‌شناسیم.



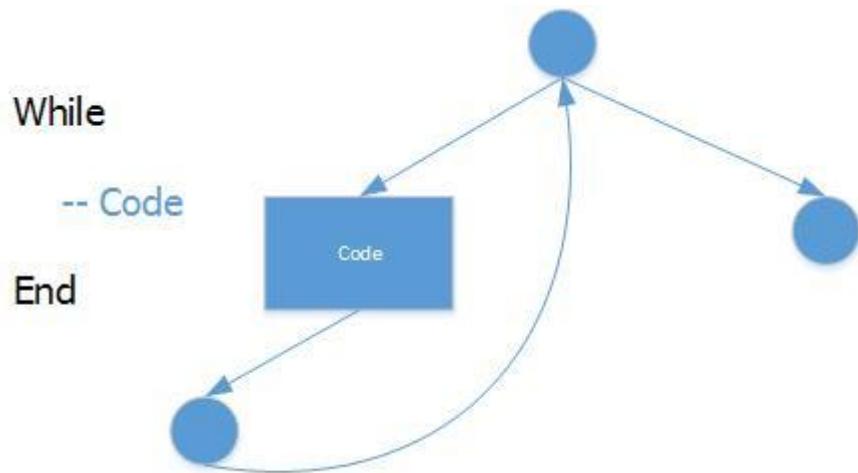
شکل ۹-۴. تبدیل عملگر شرط کامل به گراف

شکل ۱۰-۳ را با عنوان گراف عملگر شرط تو در تو می‌شناسیم که پیاده‌سازی کدهای چند شرطی و یا Switch Case به گراف پس از تبدیل به زبان استقلال الگوریتم می‌باشد.



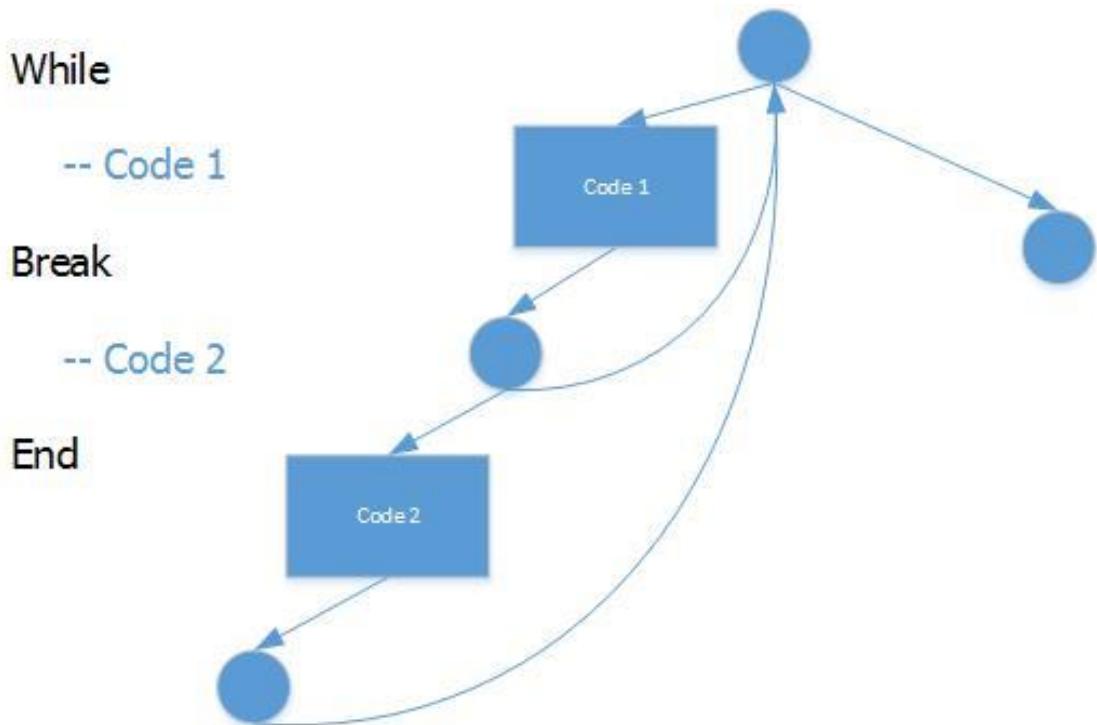
شکل ۱۰-۳. تبدیل عملگر شرط تو در تو به گراف

گراف عملگر حلقه شکل ۱۱-۳ نشان‌دهنده تبدیل ساده‌ترین نوع عملگر حلقه به گراف می‌باشد.



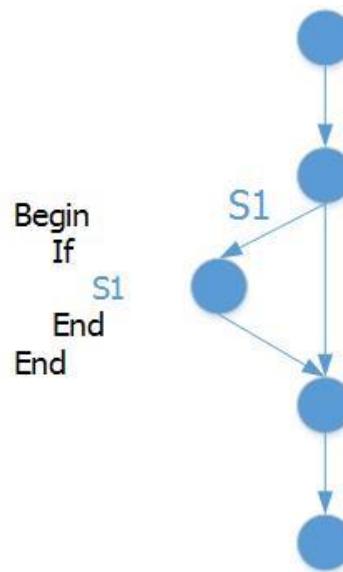
شکل ۱۱-۳. تبدیل عملگر حلقه ساده به گراف

شکل ۱۲-۳ نشان‌دهنده تبدیل عملگر حلقه با پرش به گراف می‌باشد.

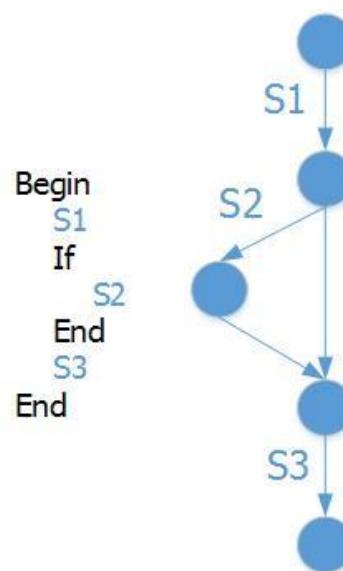


شکل ۱۲-۳. تبدیل عملگر حلقه با پرش به گراف

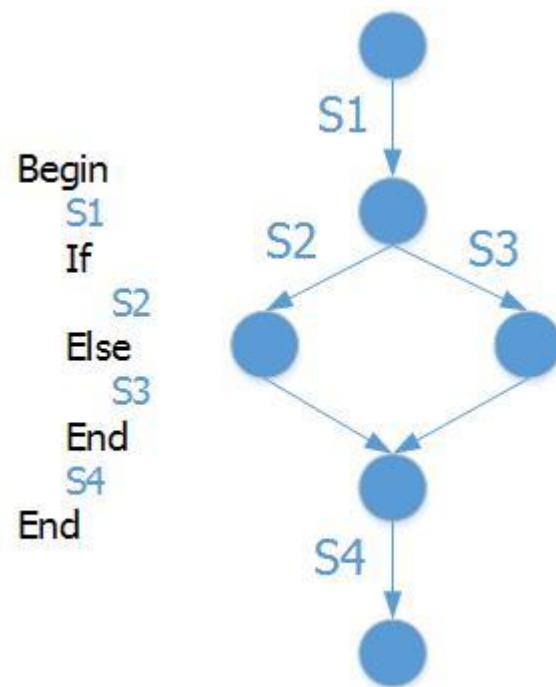
در ادامه به ذکر چند مثال پرکاربرد از کدهای تبدیل شده به گراف می‌پردازیم. همان طور که در کد این مثالها نمایش داده شده است برای وضوح بیشتر مثالها به جای استفاده از نمادهای جدول نگاشت از حرف S که به ترتیب شماره‌گذاری شده است استفاده می‌شود. بدیهی است که نمادهای به کاررفته در کد واقعی استقلال الگوریتم جایگزین این S ‌ها خواهد شد. شکل ۱۳-۳ الى ۲۲-۳ مثالهایی از کدهای پرکاربرد مطابق با قالب استاندارد کد استقلال الگوریتم می‌باشد.



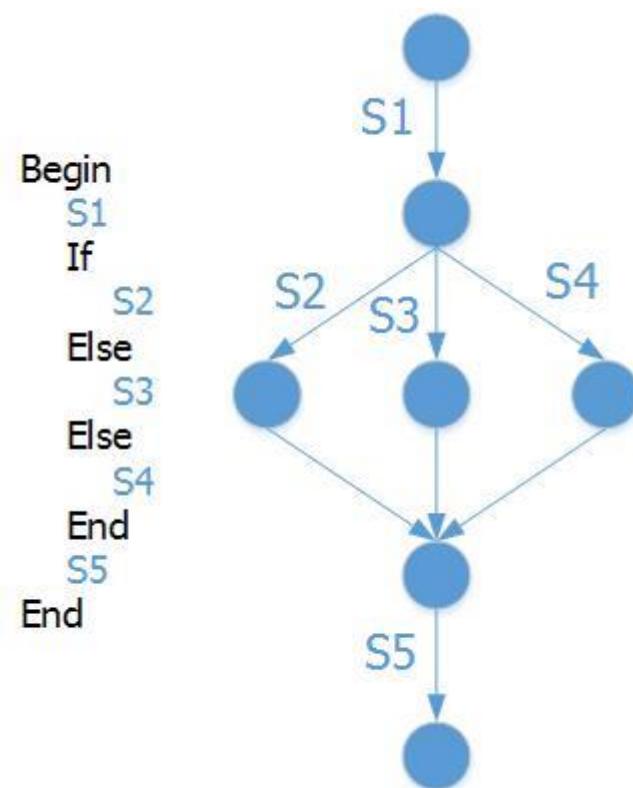
شکل ۱۳-۳. پیادهسازی شرط ساده بدون هیچ کد اضافی



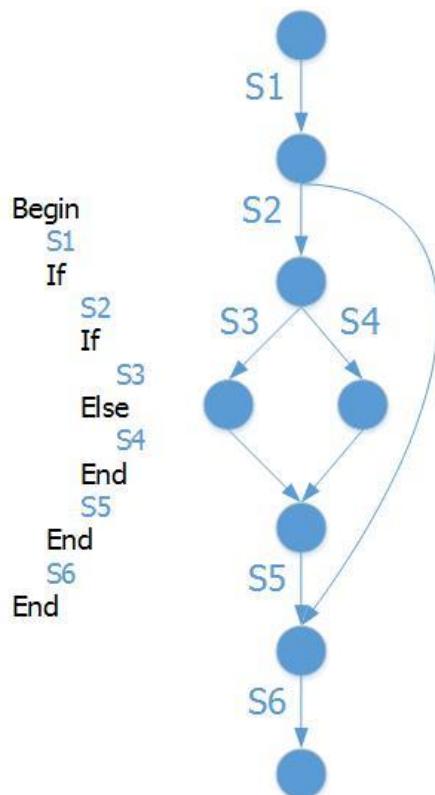
شکل ۱۴-۳. پیادهسازی شرط ساده با کدهای قبل و بعد آن



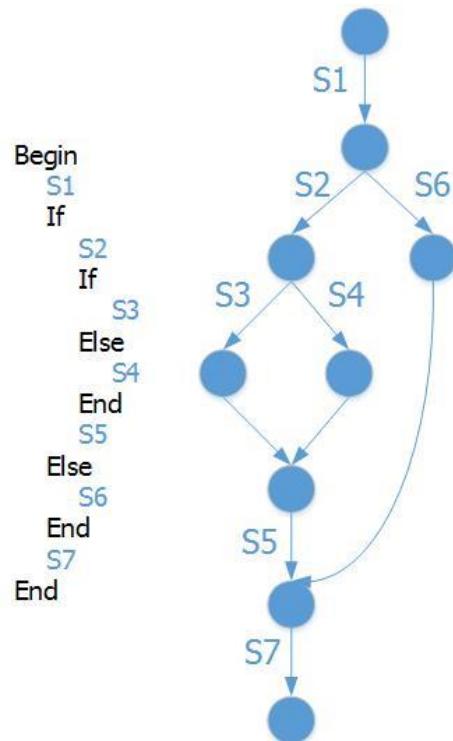
شکل ۳-۱۵. پیاده‌سازی شرط کامل



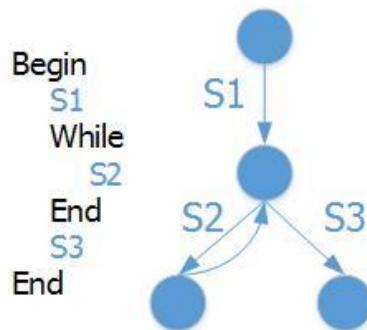
شکل ۳-۱۶. پیاده‌سازی شرط تو در تو



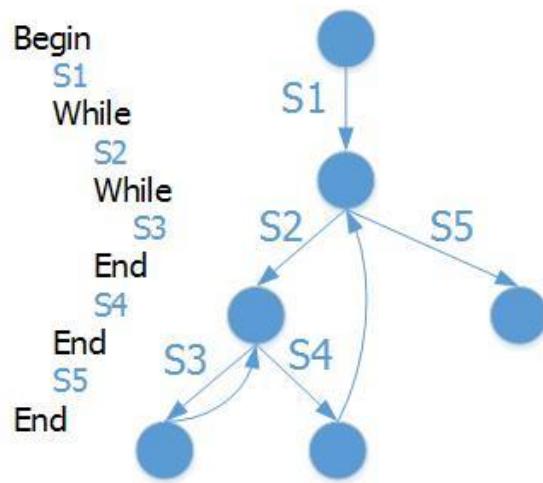
شکل ۱۷-۳. پیاده‌سازی یک شرط کامل در یک شرط ساده



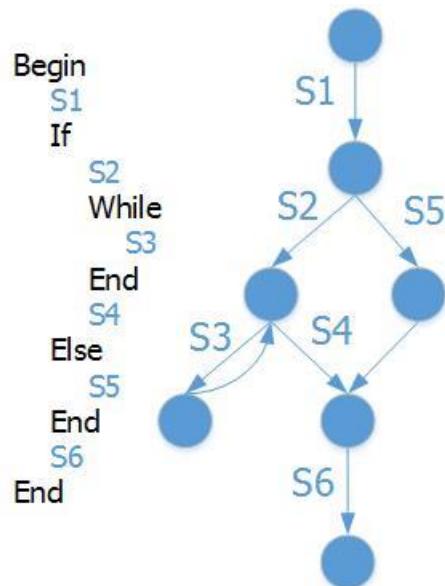
شکل ۱۸-۳. پیاده‌سازی یک شرط کامل در یک شرط کامل دیگر



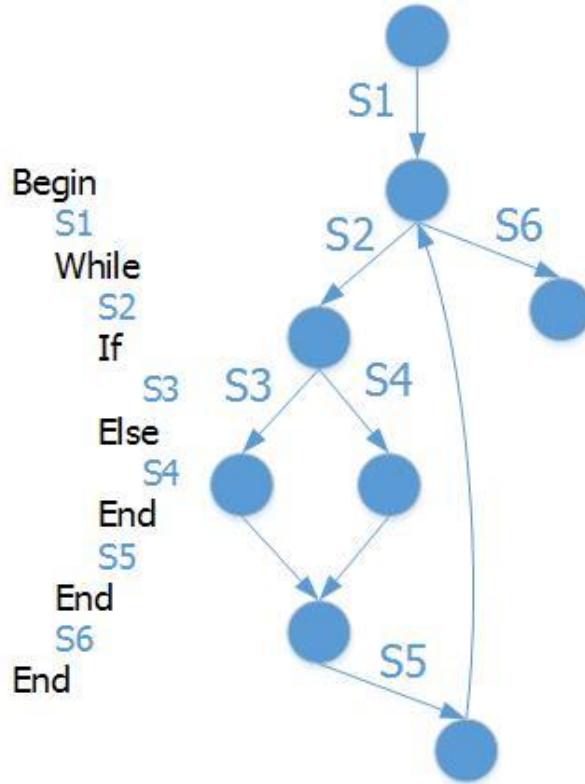
شکل ۱۹-۳. پیاده‌سازی حلقه ساده



شکل ۲۰-۳. پیاده‌سازی یک حلقه ساده داخل حلقه‌ای دیگر



شکل ۲۱-۳. پیاده‌سازی یک حلقه داخل یک شرط کامل



شکل ۲۲-۳. پیاده‌سازی یک شرط کامل داخل یک حلقه ساده

در ادامه به بررسی روش ارزیابی گراف استقلال برای محاسبه درجه استقلال دو الگوریتم می‌پردازیم.

۲-۴-۳. ارزیابی گراف استقلال الگوریتم

ابتدا برای ارزیابی درجه استقلال دو الگوریتم به معروفی روشنی جهت ذخیره‌سازی گراف آن می‌پردازیم. همان‌گونه که در مثال‌های ذکر شده مشاهده می‌شود همیشه یال‌هایی بدون وزن در گراف وجود دارند که صرفاً برای نمایش عملکرد کامل گراف به کاربرده می‌شوند و در امر ارزیابی نیاز به ذخیره آن‌ها نیست لذا برای ذخیره گراف در حافظه از ذخیره‌سازی آن‌ها صرف‌نظر خواهیم کرد تا هم حجم حافظه بهینه استفاده شود و هم سرعت اجرای مقایسه افزایش یابد. در این تحقیق برای ذخیره هر گراف در حافظه از آرایه‌ای به اندازه تمامی یال‌های وزن‌دار آن استفاده می‌شود که وزن هر یال (که می‌تواند شامل چندین خط از نمادهای جدول نگاشت استاندارد باشد) در یک خانه منحصر به فرد ذخیره می‌شود. برای ارزیابی گراف استقلال دو الگوریتم کافی است آرایه‌های آن دو الگوریتم را باهم

مقایسه کنیم. جهت ارزیابی آرایه‌های دو الگوریتم، ماتریسی با عنوان ماتریس درجه وابستگی کد^{۱۷۷} یا CDDM را همانند شکل ۲۳-۳ طوری فرض می‌کنیم که هر خانه آن مقایسه‌ای از وزن یال‌های گراف الگوریتم اول باهر کدام از وزن یال‌های گراف الگوریتم دوم باشد.

$$CDDM = \begin{bmatrix} \text{Alg 1}[1] & \text{Alg 1}[2] & \cdots & \text{Alg 1}[n] \\ \text{Alg 2}[1] & a_{11} & a_{12} & \cdots & a_{1n} \\ \text{Alg 2}[2] & a_{21} & a_{22} & \ddots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{Alg 2}[m] & a_{m1} & a_{m2} & & a_{mn} \end{bmatrix}$$

شکل ۲۳-۳. ماتریس درجه وابستگی کد

در شکل ۳ ۲۳-۳ متغیر n تعداد خانه‌های آرایه الگوریتم اول و متغیر m تعداد خانه‌های آرایه الگوریتم دوم می‌باشد. مقادیر خانه‌های ماتریس بر اساس رابطه ۷-۳ محاسبه می‌شود.

$$a_{ij} = compare(\text{Alg 1}[i], \text{Alg 2}[j]) \quad (7-3)$$

در رابطه ۷-۳ متغیرهای i و j شماره سطر و ستون ماتریس شکل ۲۳-۳ می‌باشند و تابع $compare(x, y)$ بر اساس شبه کد ۲۴-۳ محتوای دو خانه از آرایه‌ها را باهم مقایسه می‌کند.

Function Compare (Cell1, Cell2) **Return** [CDD]

Count = 0

While we have Symbol in Cell1

Sym1 = **Select** an Symbol in Cell1

Foreach Sym2 in Cell2

If Sym2 = Sym1 is found **then**

Count++

Break

End If

End Foreach

End while

MSymbol = Max-Sym (Cell1, Cell2)

Result = Count / MSymbol

شکل ۲۴-۳. شبه کد مقایسه محتوای دو خانه از آرایه‌های استقلال الگوریتم

¹⁷⁷ Code Dependence Degree Matrix

در شکل ۲۴-۳ به ازای هر نماد در هر خانه از آرایه اول به دنبال اولین نماد عیناً مشابه در خانه‌ی آرایه دوم می‌گردد و تعداد نمادهای مشابه را می‌شمارد. در نهایت جهت نرمال سازی درجه استقلال بین صفر و یک تعداد نمادهای مشابه را بر مقدار بیشینه تعداد نمادهای موجود بین خانه اول و دوم تقسیم می‌کند. همان طور که در تعریف شکل ۲۴-۳ نشان داده شده است مقدار به دست آمده از تابع $compare(x, y)$ را با عنوان درجه وابستگی هر خانه^{۱۷۸} یا همان CDD می‌شناسیم.

جهت به دست آوردن مقدار استقلال الگوریتم بر اساس ماتریس وابستگی، فرض می‌کنیم که ماتریس فعلی ماتریس مرحله اول یا همان $CDDM_0$ نام دارد. در این ماتریس خانه‌ای را پیداکرده که بیشترین مقدار را دارد که ما مقدار آن را به متغیر $MaxCell_0$ تخصیص می‌دهیم. با حذف سطر و ستونی که خانه بیشینه در آن قرار دارد ماتریس $CDDM_1$ را می‌سازیم و به طریق مشابه خانه‌ای که در آن بیشترین مقدار وجود دارد در $MaxCell_1$ نگهداری کرده و سطر و ستون خانه بیشینه را حذف می‌کنیم تا ماتریس مرحله بعد ساخته شود. این کار را به تعداد n بار می‌توان انجام داد که حداقل تعداد خانه‌های دورآرایه‌ی الگوریتم اول و دوم می‌باشد. بر اساس تعریف بالا درجه استقلال الگوریتم مطابق با رابطه ۸-۳ محاسبه می‌شود.

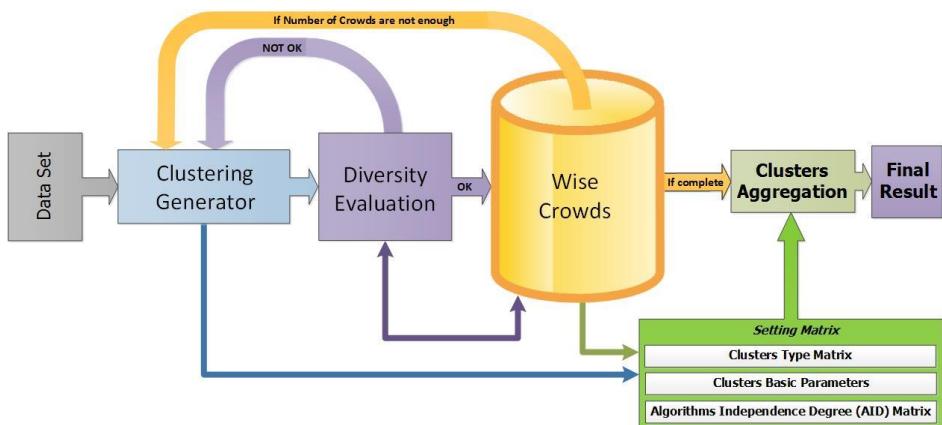
$$AID(\text{Alg 1}, \text{Alg 2}) = 1 - \frac{1}{m} \sum_{i=1}^n MaxCell_i \quad (8-3)$$

در رابطه ۸-۳ نمادهای Alg1 و Alg2 به ترتیب نشان دهنده آرایه‌های الگوریتم اول و دوم می‌باشد و متغیر m حداکثر تعداد بین خانه‌های آرایه‌های الگوریتم اول و دوم است. طبق این رابطه درجه استقلال دو الگوریتم برابر تفاضل یک از میانگین خانه‌های بیشینه در ماتریس‌های درجه وابستگی کد در مراحل بالا است که جهت نرمال سازی بین صفر و یک بر m تقسیم شده است. از این پس درجه استقلال الگوریتم^{۱۷۹} را با عنوان AID می‌شناسیم.

¹⁷⁸ Cell Dependence Degree

¹⁷⁹ Algorithms Independence Degree

٣-٤-٣. چهار چوب خوش بندی خردمند مبتنی بر گراف استقلال الگوریتم



شكل ٣-٢٥. چهار چوب خوش بندی خردمند مبتنی بر گراف استقلال الگوریتم

همان طور که پیشتر به آن اشاره شد در روش پیشنهادی دوم این مقاله استقلال الگوریتم به عنوان وزنی برای ترکیب نتایج نهایی استفاده خواهد شد لذا در شکل ۲۵-۳ بخش ارزیابی استقلال حذف شده و پس از ارزیابی پراکندگی هر الگوریتم آن را در مجمع خردمند ذخیره می‌کنیم. در اصل اینجا مجمع خردمند شامل نتایج ارزیابی شده (نسبت به پراکندگی) به همراه درجه استقلال آن نتایج (احتمال درستی آن‌ها نسبت به هم) به عنوان وزن جهت ترکیب می‌باشند. همان‌گونه که در شکل نشان داده شده است این چهارچوب نیز غیرمت مرکز بوده و تمامی قوانینی که در بخش عدم مرکز و ارزیابی پراکندگی در روش اول به آن اشاره شد در اینجا نیز صدق خواهد کرد. از این روی در این بخش تنها به تشریح روش ارزیابی استقلال بر اساس روش مبتنی بر گراف و همچنین روش ترکیب نتایج جمع‌آوری شده در مجمع خردمند به صورت وزن‌دار بسنده می‌کنیم.

٤-٣-٣-١. ارزیابی استقلال الگوریتم

مطابق شکل ۲۵-۳ نتایج ارزیابی درجه استقلال الگوریتم‌ها در ماتریس AIDM نگهداری می‌شود. اندازه این ماتریس $n \times n$ است که n تعداد الگوریتم‌هایی هستند که قرار است در خوشبندی شرکت کنند. روش محاسبه سطر و ستون این ماتریس طبق رابطه ۹-۳ محاسبه می‌شود:

$$a_{ij} = \begin{cases} AID(A \lg 1, A \lg 2) & i \neq j \\ -1 & i = j \end{cases} \quad (9-3)$$

رابطه ۹-۳ توضیح می‌دهد که اگر دو الگوریتم غیر هم نام باشد مطابق با رابطه ۸-۳ درجه استقلال آن‌ها محاسبه می‌شود، و اگر دو الگوریتم هم نام باشد چون باید همانند روش اول با در نظر گرفتن پارامترهای اساسی آن دو الگوریتم مطابق شبه کد شکل ۲-۳ به محاسبه درجه استقلال بپردازیم درجه استقلال آن را در این ماتریس منفی یک در نظر می‌گیریم تا پس از تولید پارامترهای اساسی حین اجرای الگوریتم به محاسبه آن بپردازیم. در این حالت استقلال الگوریتم هم در سطح عملکرد الگوریتم و هم در سطح پارامترهای اساسی الگوریتم در نظر گرفته می‌شود. در شکل ۲۵-۳ پس از اتمام کار انتخاب خوشها و تشکیل جامعه خردمند، درجه استقلال نهایی برای هر الگوریتم خوشبندی برابر با میانگین مجموع درجه‌های استقلال آن الگوریتم نسبت به الگوریتم‌های غیر هم نام و پارامترهای اساسی الگوریتم‌های هم نام می‌باشد. رابطه ۱۰-۳ روش محاسبه درجه استقلال نهایی هر الگوریتم را در حین اجرا نشان می‌دهد:

$$\forall \text{Alg}_i \in \text{WisedCrowd} \Rightarrow AI = \frac{1}{n+m} \left(\sum_n AIDM[i,n] + \sum_m BPI[i,m] \right) \quad (10-3)$$

در این رابطه Alg_i نشان‌دهنده الگوریتمی است که قرار است درجه استقلال نهایی آن محاسبه شود و WisedCrowd جامعه خردمند و یا همان نتایج اولیه انتخاب شده می‌باشد و n تعداد الگوریتم‌های غیر هم نام و m تعداد الگوریتم‌های هم نام با این الگوریتم در جامعه خردمند می‌باشد. در این رابطه به ازای هر الگوریتم غیر هم نام درجه استقلال آن نسبت به الگوریتم از ماتریس $AIDM$ خواند می‌شود و به ازای هر الگوریتم هم نام درجه استقلال آن نسبت به الگوریتم بر اساس پارامترهای اساسی الگوریتم که حین اجرا تولید می‌شوند با استفاده ازتابع BPI محاسبه می‌شود. بدیهی است که AI همواره مقداری بین صفر و یک خواهد داشت چون از میانگین اعدادی که همیشه بین صفر و یک هستند به دست می‌آید. همچنین جهت استفاده از درجه استقلال نهایی به دست آمده به عنوان وزن در ترکیب نتایج اولیه آن‌ها را در ماتریسی با عنوان ماتریس استقلال الگوریتم‌ها^{۱۸۰} یا همان AIM نگهداری می‌کنیم. قابل ذکر است که این ماتریس فقط و فقط استقلال نهایی الگوریتم‌هایی که موفق به ورود به جامعه خردمند شده‌اند را نگهداری می‌کند و نه تمام الگوریتم‌هایی که در تشکیل نتایج اولیه خوشبندی ترکیبی نقش داشته‌اند.

¹⁸⁰ Algorithms Independence Matrix

۲-۳-۴-۳. روش انباشت مدارک وزن دار

در سال های اخیر بسیاری از مقالات جهت ترکیب نتایج خوشه بندی از روش انباشت مدارک به خاطر کالایی بالای آن استفاده کرده اند [8, 9, 19, 21, 67]. در این روش برای ترکیب نتایج از رابطه ۵۶-۲ استفاده می شود. همان طور که پیشتر ذکر شد در این رابطه پارامتر $n_{i,j}$ تعداد دفعاتی است که جفت نمونه های i و j باهم در یک خوشه گروه بندی شده اند و همچنین $m_{i,j}$ تعداد نمونه برداری هایی است که هر دوی این جفت نمونه ها به طور همزمان در آن ظاهر شده اند. با توجه به نحوه محاسبه پارامتر $n_{i,j}$ می توان گفت شمارش تعداد نمونه ها یعنی هر نتیجه با وزن مساوی و برابر با یک در جواب نهایی شرکت می کند. در این تحقیق دیدگاه جدیدی در مورد رابطه ۵۶-۲ مطرح خواهد شد که در آن قرار است درجه نهایی استقلال هر الگوریتم به عنوان وزنی برای احتمال درستی هر نتیجه در نظر گرفته شود. از این روی از رابطه ۱۱-۳ که ما آن را با عنوان روش انباشت مدارک وزن دار^{۱۸۱} یا همان WEAC نام گذاری کرده ایم برای ترکیب نتایج استفاده خواهیم کرد.

$$C(i,j) = \frac{\sum_{n_{i,j}} res(i) \times AIM[Alg_i]}{m_{i,j}} \quad (11-3)$$

در رابطه ۱۱-۳ متغیر $AIDM$ نشان دهنده ماتریس درجه استقلال الگوریتم می باشد و $n_{i,j}$ تعداد دفعاتی است که جفت نمونه های i و j باهم در یک خوشه گروه بندی شده اند وتابع res مقدار مربوط به نمونه i -ام (که این مقدار برابر با مقدار نمونه j -ام نیز است) را از نتایج اولیه خوشه بندی بر می گرداند و Alg_i شماره ای الگوریتمی که دو نمونه i و j را تولید کرده است بر می گرداند که متناسب با آن درجه استقلال نهایی آن الگوریتم از ماتریس AIM خوانده خواهد شد. همچنین در این رابطه $m_{i,j}$ تعداد نمونه برداری هایی است که هر دوی این جفت نمونه ها به طور همزمان در آن ظاهر شده اند.

¹⁸¹ Weighted Evidence Accumulation Clustering

۳-۴-۳. شبه کد خوشبندی خردمند مبتنی بر گراف استقلال الگوریتم

همان طور که پیش تر اشاره شد شکل ۲۵-۳ چهارچوب روش پیشنهادی دوم این تحقیق را ارائه می دهد. این فرآیند با تولید نتایج اولیه شروع می شود و در ادامه با ارزیابی پراکندگی و وزن های به دست آمده از ماتریس استقلال الگوریتم اقدام به تولید جامعه خردمند می کنیم. در نهایت با استفاده از روش انباشت مدارک وزن دار افزایش ای جمع آوری شده در جامعه خردمند با یکدیگر ادغام شده و نتیجه نهایی را تولید می کند. شکل ۲۶-۳ نشان دهنده شبه کد روش پیشنهادی دوم است.

Function WCboAIG (Dataset, Kb, dT, cT) **Return** [Result, nCrowd]

Initialized nCrowd to zero

While we have base cluster

[IDX, Basic-Parameter] = Generate-Basic-Algorithm (Dataset, Kb*cT)

If (Diversity (IDX) > dT) **then**

 Find the Algoritms AID from AIDM

 Insert idx, AID, and Basic-Parameter to Crowd-Partitions

 Crowd = Crowd + 1

End if

End while

Generate AIM matrix

W-Co-Acc = WEAC (Crowd-Partition, AIM)

Z = Average-Linkage (W-Co-Acc)

Result = Cluster (Z, Kb)

شکل ۲۶-۳. شبه کد خوشبندی خردمند مبتنی بر گراف استقلال الگوریتم

در این شکل Kb تعداد خوشها در الگوریتم پایه می باشد. همانند روش اول پارامترهای dT و cT به ترتیب مقادیر آستانه برای ارزیابی پراکندگی و عدم تمرکز هستند.تابع Generate-Basic-Algorithm نتایج اولیه (افزارهای) را با استفاده از الگوریتمهای خوشبندی های پایه تولید می کند. تابع Diversity برای ارزیابی پراکندگی به کار می رود و تابع WEAC ماتریس همبستگی را برای تولید نتیجه نهایی با استفاده از نتایج اولیه به صورت روش انباشت مدارک وزن دار بر اساس رابطه ۱۱-۳ تولید می کند. برای تولید دندوگرام از ماتریس همبستگی ما از الگوریتم پیوندی میانگین استفاده کرده ایم چون نتایج

تجربی این تحقیق نشان داده است که این روش بهترین دقیقت را داراست. در اینجا تابع Average-Linkage نشان‌دهنده الگوریتم پیوندی میانگین است. در نهایت تابع Cluster بر اساس تعداد خوش تعبیین شده نتیجه نهایی را از روی دندوگرام تشکیل می‌دهد.

به عنوان نکته پایانی می‌توان به این موضوع اشاره کرد که در شبه کد شکل ۲۶-۳ با استفاده از روش ارزیابی استقلال مبتنی بر گراف و به کارگیری آن به عنوان وزن در روش انباشت مدارک وزن‌دار ما میزان تأثیر رأی هر الگوریتم را با تغییر اندازه در سطح‌های دندوگرام بر روی نتیجه نهایی اعمال می‌کنیم. به عنوان مثال می‌توان گفت اگر دو الگوریتم با درجه استقلال پایین در تولید نتیجه‌ی شبه کد شکل ۲۶-۳ شرکت کنند و نتایج مشابه داشته باشند آنگاه روش پیشنهادی دوم فقط و فقط به اندازه میزان استقلال آن دو الگوریتم شکل دندوگرام را تغییر می‌دهد که بسیار کمتر از وقتی است که دو الگوریتم کاملاً مستقل (یعنی درجه استقلال آن‌ها برابر با یک باشد) با نتایج برابر در تشکیل نتیجه نهایی شرکت می‌کنند.

فصل چهارم

پیاده‌سازی و

تحلیل نتایج

۴. پیاده‌سازی و تحلیل نتایج

۱-۱. مقدمه

در این فصل نتایج آزمایش‌های تجربی این تحقیق را جهت ارزیابی الگوریتم‌های پیشنهادی ارائه خواهیم کرد. از این روی در ادامه در بخش مجموعه داده، ابتدا به بررسی داده‌های استاندارد به کاررفته در این تحقیق خواهیم پرداخت. پس از معرفی داده‌ها و مشخصات آنها، در بخش مدل‌سازی الگوریتم‌ها به زبان استقلال الگوریتم لیستی از الگوریتم‌های پایه که در ساخت نتایج اولیه خوشبندی از آنها استفاده شده است ارائه می‌گردد و همچنین پیاده‌سازی کدهای الگوریتم‌های ذیل به زبان استاندارد استقلال الگوریتم که پیشتر به آن اشاره شد نیز ارائه خواهد شد. در بخش ابزار تحلیلگر کد استقلال الگوریتم^{۱۸۲} به معرفی نرم‌افزاری که مناسب با استانداردهای این تحقیق برای تبدیل خودکار کد استقلال الگوریتم به گراف و ارزیابی آن به زبان برنامه‌نویسی C# در مجموعه Microsoft Visual Studio 2012 طراحی و ساخته شده است می‌پردازیم. سرانجام، در بخش نتایج آزمایش‌ها دقت و میزان NMI نتایج نهایی الگوریتم‌های پیشنهادی این تحقیق نسبت به کلاس‌های واقعی داده را با روش‌های پیشین مقایسه می‌کنیم و همچنین تأثیر پارامترهای معرفی شده در این تحقیق همچون پراکنده‌گی، استقلال و عدم تمرکز بر روی کارایی نتایج و زمان اجرای الگوریتم‌ها را بررسی خواهیم کرد. کلیه نتایج ارائه شده در این بخش توسط پیاده‌سازی و شبیه‌سازی الگوریتم‌ها در نرم‌افزار Matlab R2013a (8.1.0.604) تولید و ارائه شده‌اند.

۲-۱. مجموعه داده

در این تحقیق نتایج تجربی آزمایش‌ها بر روی چهارده مجموعه داده استاندارد برای ارزیابی روش پیشنهادی گزارش شده‌اند. بیشتر مجموعه داده‌ها در این تحقیق از مجموعه داده‌های استاندارد UCI [76] می‌باشند که تقریباً نتایج تمام مطالعات اخیر دنیا در زمینه خوشبندی با استفاده از این مجموعه داده‌ها گزارش می‌شوند. علاوه بر آن از داده Halfring که در کارهای تحقیقاتی عظیمی و همکاران [1, 2, 4, 5, 6, 7] و علیزاده و همکاران [67, 8, 9] به عنوان یک داده مصنوعی با شکل غیر

¹⁸² CAIL code analyzer

کروی که تشخیص آن توسط الگوریتم‌های خوشبندی پایه سخت می‌باشد نیز مورد استفاده قرار گرفته است. جدول ۱-۴ مشخصات مجموعه داده به کاررفته در ارزیابی الگوریتم‌های این تحقیق را نشان می‌دهد.

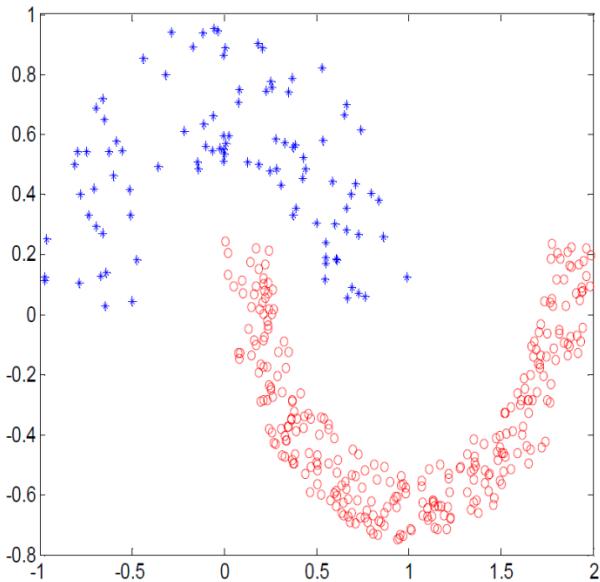
جدول ۱-۴. مجموعه داده

No.	Name	Feature	Class	Sample
1	Half Ring	2	2	400
2	Iris	4	3	150
3	Balance Scale	4	3	625
4	Breast Cancer	9	2	683
5	Bupa	6	2	345
6	Galaxy	4	7	323
7	Glass	9	6	214
8	Ionosphere	34	2	351
9	SA Heart	9	2	462
10	Wine	13	2	178
11	Yeast	8	10	1484
12	Pendigits	16	10	10992
13	Statlog	36	7	6435
14	Optdigits	62	10	5620

در انتخاب مجموعه داده جدول ۱-۴ سعی شده است که از هر سه نوع داده کوچک، متوسط و بزرگ چند نمونه داده انتخاب شود تا تأثیر اندازه‌ی داده بر روی روش عملکرد الگوریتم و نتایج آن کاهش یابد. ویژگی‌های مجموعه داده به کاررفته در این تحقیق جهت حذف تأثیر مقیاس ابعاد داده بر روی نتایج با میانگین صفر و وردایی^{۱۸۳} یک، $N(0,1)$ نرمال شده‌اند. از آنجایی که به جز داده Halfring مشخصات و روش تهیه تمامی داده‌های جدول ۱-۴ توسط نیومن و همکاران به خوبی در [76]

¹⁸³ Variance

توضیح داده شده است در ادامه به ارائه شکل Halfring و تشریح مشخصات آن بسنده می‌کنیم. مطابق با شکل ۱-۴ داده Halfring در نمودار دو بعدی اعداد نشان داده شده است.



شکل ۱-۴. مجموعه داده [1] Halfring

مجموعه داده Halfring که یک مجموعه داده مصنوعی است همان طور که در شکل ۱-۴ نشان داده شده است شامل دو خوشی که در اینجا با رنگ‌های آبی و قرمز آنها را جدا کرده‌ایم می‌شود. تعداد داده‌ها در دو خوشی با ۱۰۰ نقطه و ۳۰۰ نقطه، نامتوازن است و در مرکز شکل هندسی این مجموعه داده، دو کلاس داده در بخشی از فضای باهم در یک راستا همپوشانی دارد. از این روی الگوریتم‌هایی همچون K-means که بر اساس شکل ابر کره کار می‌کنند به تنها یک قادر به تشخیص دو خوشی نیستند پس Halfring یکی از مناسب‌ترین و سخت‌ترین نوع از مجموعه داده‌ی دو بعدی برای ارزیابی هر الگوریتم خوشبندی می‌باشد.

۳-۴. مدل‌سازی الگوریتم‌ها به زبان استقلال الگوریتم

جدول ۲-۴ لیست الگوریتم‌هایی است که در این تحقیق برای ساخت نتایج اولیه خوشبندی استفاده شده‌اند. در تهیه این لیست سعی شده تا یک یا چند الگوریتم از هر یک از انواع نامبرده شده در دسته‌بندی شکل ۵-۳ بکار گرفته شود. در این انتخاب آن دسته الگوریتم‌هایی که با هر بار اجرا یک جواب بر اساس یک متغیر تصادفی تولید می‌کنند همانند K-means یک الگوریتم انتخاب شود. همچنین سعی شده است که با تغییر در معیارهای اندازه‌گیری و یا به کارگیری انواع آن دسته از

الگوریتم‌هایی که همیشه روی داده یک جواب منحصر به فرد را تولید می‌کند همانند Linkage پراکنده‌گی در نتایج اولیه را به حداقل برسانیم.

جدول ۲-۴. لیست مجموعه الگوریتم‌های پایه

No.	Algorithm Name	Code
1	K-Means	K
2	Fuzzy C-Means	F
3	Median K-Flats	M
4	Gaussian Mixture	G
5	Subtractive Clustering	SUB
6	Single-Linkage Euclidean	SLE
7	Single-Linkage Hamming	SLH
8	Single-Linkage Cosine	SLC
9	Average-Linkage Euclidean	ALE
10	Average-Linkage Hamming	ALH
11	Average-Linkage Cosine	ALC
12	Complete-Linkage Euclidean	CLE
13	Complete-Linkage Hamming	CLH
14	Complete-Linkage Cosine	CLC
15	Ward-Linkage Euclidean	WLE
16	Ward-Linkage Hamming	WLH
17	Ward-Linkage Cosine	WLC
18	Spectral clustering using a sparse similarity matrix	SPS
19	Spectral clustering using Nystrom method with orthogonalization	SPN
20	Spectral clustering using Nystrom method without orthogonalization	SPW

برای اجرای الگوریتم روش پیشنهادی دوم جهت تولید نتایج اولیه و ارزیابی آن‌ها ابتدا باید الگوریتم‌های پایه جدول ۲-۴ را مدل‌سازی کرد و استقلال این الگوریتم‌ها را ارزیابی کنیم. از این روی قبل از ارائه نتایج آزمایش‌ها تجربی این تحقیق به ارائه جدول نگاشت استاندارد کد، کدهای استقلال هر الگوریتم و ماتریس AIDM در این بخش و معرفی نرم‌افزاری ساخته شده برای این تحقیق جهت تبدیل کد استقلال به گراف و ارزیابی آن به صورت خودکار در بخش بعد می‌پردازد. جدول ۳-۴ نگاشت استاندارد کدهای به کاررفته در این تحقیق را نمایش می‌دهد.

جدول ۴-۳ جدول نگاشت استاندارد کد

No.	Symbol	Description
1	R(1)	Generate x random number
2	R(2)	Random Selection
3	M(1)	$Y = \text{EuclidianDistance}(A, B)$
4	M(2)	$c_j = \frac{\sum_{i=1}^N u_{ij}^m \times x_i}{\sum_{i=1}^N u_{ij}^m}$
5	M(3)	$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\ x_i - c_j\ }{\ x_i - c_k\ } \right)^{\frac{2}{m-1}}}$
6	M(4)	Do exp function: $S = e^{(-1 \times (A^2 / 2 \times \sigma^2))}$
7	M(5)	Do laplacian function: $L = D^{-\frac{1}{2}} \times S \times D^{-\frac{1}{2}}$
8	M(6)	Largest magnitude
9	M(7)	Smallest magnitude
10	M(8)	Normalizing A and B
11	M(9)	$Y = \text{HammingDistance}(A, B)$
12	M(10)	$Y = \text{CosinDistance}(A, B)$
13	M(11)	$D_{SL}(C_i, C_j) = \min_{a \in C_i, b \in C_j} d(a, b)$
14	M(12)	$D_{cl}(C_i, C_j) = \max_{a \in C_i, b \in C_j} d(i, j)$
15	M(13)	$D_{AL}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{a \in n_i, b \in n_j} d(a, b)$
16	M(14)	$D_{WL}(C_i, C_j) = \sqrt{\frac{2n_i n_j}{(n_i, n_j)} \ \bar{a} - \bar{b}\ _2}$
17	M(15)	$Q(\theta', \hat{\theta}^{(j)}) = E(\ell_0(\theta'; t) z, \hat{\theta}^{(j)})$
18	M(16)	$(\ell^*, m^*) = \arg \min_{(\ell, m)} d(\ell, m)$
19	M(17)	$Z = X/Y$
20	M(18)	$d_x P_i = - \frac{(P_i x x^T - P_i x x^T P_i^T P_i)}{\sqrt{1 - \ P_i x\ ^2}}$
21	M(19)	$i^* = \arg \max_{1 \leq i \leq K} \ P_i x\ $
22	F(1)	Assign each object to closest centroid/subspace
23	F(2)	Generate (t-nearest-neighbor) sparse distance matrix

24	F(3)	Convert distance matrix to similarity matrix
25	F(4)	Do orthogonalization
26	F(5)	Restore cluster labels in original order
27	F(6)	Compute the proximity matrix
28	F(7)	Merge two closest cluster
29	F(8)	$Y = \text{Subclass}(X)$
30	F(9)	Update $P_{i^*} : P_{i^*} \mapsto P_{i^*} - dtd_{x^*} P_{i^*}$

در جدول بالا سه گروه اعداد تصادفی (R)، معادلات ریاضی (M) و توابع (F) وجود دارند. در تهیه این جدول بیشتر سعی شده تا از شبیه کدهای نمادگذاری کدهای الگوریتمها استفاده شود تا هر چه بیشتر کد از جزئیات برنامه‌نویسی که نقش خاصی در استقلال ندارند به دور بماند. شکل ۱-۴ کد الگوریتم K-means به زبان استقلال الگوریتم بر اساس شبیه کد شکل ۹-۲ می‌باشد.

```

Begin
R(1)
While
    F(1)
    M(1)
End
End

```

شکل ۲-۴. الگوریتم K-means

شکل ۳-۴ توصیف الگوریتم FCM به زبان استقلال الگوریتم بر اساس شبیه کد شکل ۱۰-۲ می‌باشد.

```

Begin
R(1)
While
    M(2)
    M(3)
End
End

```

شکل ۳-۴. الگوریتم FCM

شكل ٤-٤ توصيف الگوریتم Median K-Flats به زبان استقلال الگوریتم بر اساس [77] و شبه کد شکل ۱۲-۲ می‌باشد.

```
Begin
While
    R(1)
    M(19)
    M(18)
    F(9)
    F(4)
End
F(1)
End
```

شكل ٤-٤. الگوریتم Median K-Flats

شكل ٤-٥ توصیف الگوریتم Gaussian Mixture به زبان استقلال الگوریتم بر اساس [83] و توضیحات بخش ۶-۲-۱-۲-۲ و با الهام از روش پیاده‌سازی الگوریتم EM در شکل ۱۷-۲ می‌باشد.

```
Begin
While
    While
        M(15)
        M(12)
    End
    M(16)
End
M(11)
End
```

شكل ٤-٥. الگوریتم Gaussian Mixture

```
Begin
M(11)
M(12)
M(19)
F(8)
R(1)
While
    F(1)
    M(1)
End
End
```

شكل ٤-٦. الگوریتم خوشبندی Subtractive

شکل ۴-۶ توصیف الگوریتم خوشبندی Subtractive به زبان استقلال الگوریتم بر اساس [79] و شبه کد شکل ۱۱-۲ می باشد. همچنین در شکل های ۷-۴ الی ۱۸-۴ توصیف الگوریتم های سری Hamming بر اساس توضیحات بخش ۲-۱-۲ و با استفاده از معیارهای فاصله اقلیدسی، پیوندی و Cosine نشان داده شده است.

```
Begin
F(6)
M(1)
While
    M(11)
    F(7)
End
End
```

شکل ۴-۷. الگوریتم پیوندی منفرد با استفاده از معیار فاصله اقلیدسی

```
Begin
F(6)
M(9)
While
    M(11)
    F(7)
End
End
```

شکل ۴-۸. الگوریتم پیوندی منفرد با استفاده از معیار فاصله Hamming

```
Begin
F(6)
M(10)
While
    M(11)
    F(7)
End
End
```

شکل ۴-۹. الگوریتم پیوندی منفرد با استفاده از معیار فاصله Cosine

```
Begin
F(6)
M(1)
While
    M(12)
    F(7)
End
End
```

شکل ۱۰-۴. الگوریتم پیوندی کامل با استفاده از معیار فاصله اقلیدسی

```
Begin
F(6)
M(9)
While
    M(12)
    F(7)
End
End
```

شکل ۱۱-۴. الگوریتم پیوندی کامل با استفاده از معیار فاصله Hamming

```
Begin
F(6)
M(10)
While
    M(12)
    F(7)
End
End
```

شکل ۱۲-۴. الگوریتم پیوندی کامل با استفاده از معیار فاصله Cosine

```
Begin
F(6)
M(1)
While
    M(13)
    F(7)
End
End
```

شکل ۱۳-۴. الگوریتم پیوندی میانگین با استفاده از معیار فاصله اقلیدسی

```
Begin
F(6)
M(9)
While
    M(13)
    F(7)
End
End
```

شکل ۱۴-۴. الگوریتم پیوندی میانگین با استفاده از معیار فاصله Hamming

```
Begin
F(6)
M(10)
While
    M(13)
    F(7)
End
End
```

شکل ۱۵-۴. الگوریتم پیوندی میانگین با استفاده از معیار فاصله Cosine

```
Begin
F(6)
M(1)
While
    M(14)
    F(7)
End
End
```

شکل ۱۶-۴. الگوریتم پیوندی بخشی با استفاده از معیار فاصله اقلیدسی

```
Begin
F(6)
M(9)
While
    M(14)
    F(7)
End
End
```

شکل ۱۷-۴. الگوریتم پیوندی بخشی با استفاده از معیار فاصله Hamming

```
Begin
F(6)
M(10)
While
    M(14)
    F(7)
End
End
```

شکل ۴-۱۸. الگوریتم پیوندی بخشی با استفاده از معیار فاصله Cosine

شکل ۴-۱۹ الی ۲۱-۴ توصیف الگوریتم‌های سری طیفی به زبان استقلال الگوریتم بر اساس توضیحات بخش ۴-۲-۱-۲-۲ و [80, 81, 82] می‌باشد.

```
Begin
F(2)
F(3)
M(4)
M(5)
If
    M(6)
Else
    M(7)
End
R(1)
While
    F(1)
    M(1)
End
End
```

شکل ۴-۱۹. طیفی با استفاده از ماتریس شباهت نامتراکم^{۱۸۴}

¹⁸⁴ Spectral Clustering using a sparse similarity matrix

```
Begin
```

```
  F(2)
```

```
  R(2)
```

```
  M(1)
```

```
  M(1)
```

```
  F(3)
```

```
  M(8)
```

```
  F(4)
```

```
  If
```

```
    M(6)
```

```
  Else
```

```
    M(7)
```

```
 End
```

```
 R(1)
```

```
 While
```

```
   F(1)
```

```
   M(1)
```

```
 End
```

```
 F(5)
```

```
 End
```

شکل ۴-۲۰. طیفی با استفاده از روش نیستروم با متعادل ساز^{۱۸۵}

```
Begin
```

```
  F(2)
```

```
  R(2)
```

```
  M(1)
```

```
  M(1)
```

```
  F(3)
```

```
  M(8)
```

```
  If
```

```
    M(6)
```

```
  Else
```

```
    M(7)
```

```
 End
```

```
 R(1)
```

```
 While
```

```
   F(1)
```

```
   M(1)
```

```
 End
```

```
 F(5)
```

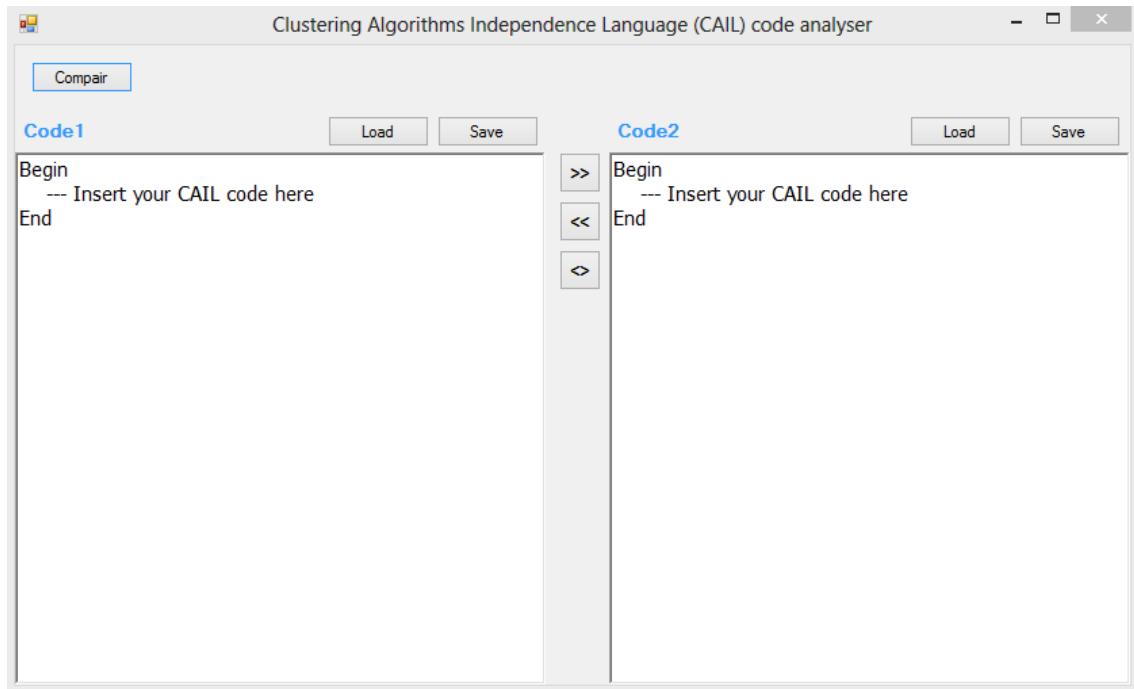
```
 End
```

شکل ۴-۲۱. طیفی با استفاده از روش نیستروم بدون متعادل ساز^{۱۸۶}

¹⁸⁵ Spectral clustering using Nystrom method with orthogonalization

¹⁸⁶ Spectral clustering using Nystrom method without orthogonalization

٤-٤. ابزار تحلیلگر کد استقلال الگوریتم



شکل ٤-٢٢. نرم افزار تحلیلگر کد استقلال الگوریتم

در شکل بالا نمایی از نرم افزار تحلیلگر کد استقلال الگوریتم به تصویر کشیده شده است. این نرم افزار که به زبان C# در محیط Microsoft Visual Studio 2012 طراحی شده است، به صورت خودکار در ابتدا کد استقلال الگوریتم را به گراف تبدیل کرده و سپس بر اساس آن وزن یالها را متناسب با تعاریف و فرضیات بخش ٣-٤-٣ در آرایه استقلال الگوریتم ذخیره می‌کند و سپس آن را با آرایه‌ای که برای الگوریتم دیگر به روش مشابه تهیه شده مقایسه می‌کند. در پایان این فرآیند درجه استقلال دو الگوریتم نسبت به هم محاسبه شده و نمایش داده می‌شود. جهت ورود اطلاعات به این نرم افزار می‌توان کدهای استقلال را در آن مستقیماً تایپ کرد یا هر فایل با قالب متن ساده را به عنوان ورودی در آن بارگذاری نمود. همچنین می‌توان کدهای هر یک از ورودی‌ها را با پسوند .cail که به صورت قراردادی پسوند پیش‌فرض این نرم افزار می‌باشد ذخیره کرد. قابل ذکر است که این نرم افزار می‌تواند درجه استقلال هر کدی با استانداردهای تعریف شده در این تحقیق را مستقل از جدول نگاشت که فرض شده برای آن محاسبه کند (فقط کدهای مقایسه شده باید بر اساس یک جدول واحد نوشته شده باشند). در شکل ٤-٣ درجه استقلال ارزیابی شده هر یک از کدهای تعریف شده در بخش قبل در ماتریس AIDM جهت استفاده در این تحقیق نشان داده شده است.

	K	F	M	G	SUB	SLE	SLH	SLC	ALE	ALH	ALC	CLE	CLH	CLC	WLE	WLH	WLC	SPS	SPN	SPW
K	-1	0.5	0.65	1	0.4	0.75	1	1	0.75	1	1	0.75	1	1	0.75	1	1	0.6	0.67	0.67
F	0.5	-1	0.9	1	0.9	1	1	1	1	1	1	1	1	1	1	1	1	0.8	0.84	0.84
M	0.65	0.9	-1	1	0.55	1	1	1	1	1	1	1	1	1	1	1	1	0.86	0.88	0.88
G	1	1	1	-1	0.94	0.84	0.84	0.84	1	1	1	0.84	0.84	0.84	1	1	1	1	1	1
SUB	0.4	0.9	0.55	0.94	-1	0.65	0.9	0.9	0.75	1	1	0.65	0.9	0.9	0.75	1	1	0.76	0.8	0.8
SLE	0.75	1	1	0.84	0.65	-1	0.25	0.25	0.25	0.5	0.5	0.25	0.5	0.5	0.25	0.5	0.5	0.9	0.92	0.92
SLH	1	1	1	0.84	0.9	0.25	-1	0.25	0.5	0.25	0.5	0.5	0.25	0.5	0.5	0.25	0.5	1	1	1
SLC	1	1	1	0.84	0.9	0.25	0.25	-1	0.5	0.5	0.25	0.5	0.5	0.25	0.5	0.5	0.25	1	1	1
ALE	0.75	1	1	1	0.75	0.25	0.5	0.5	-1	0.5	0.5	0.25	0.5	0.5	0.25	0.5	0.5	0.9	0.92	0.92
ALH	1	1	1	1	1	0.5	0.25	0.5	0.5	-1	0.5	0.5	0.25	0.5	0.5	0.25	0.5	1	1	1
ALC	1	1	1	1	1	0.5	0.5	0.25	0.5	0.5	-1	0.5	0.5	0.25	0.5	0.5	0.25	1	1	1
CLE	0.75	1	1	0.84	0.65	0.25	0.5	0.5	0.25	0.5	0.5	-1	0.5	0.5	0.25	0.5	0.5	0.9	0.92	0.92
CLH	1	1	1	0.84	0.9	0.5	0.25	0.5	0.5	0.25	0.5	0.5	-1	0.5	0.5	0.25	0.5	1	1	1
CLC	1	1	1	0.84	0.9	0.5	0.5	0.25	0.5	0.5	0.25	0.5	0.5	-1	0.5	0.5	0.25	1	1	1
WLE	0.75	1	1	1	0.75	0.25	0.5	0.5	0.25	0.5	0.5	0.25	0.5	0.5	-1	0.5	0.5	0.9	0.92	0.92
WLH	1	1	1	1	1	0.5	0.25	0.5	0.5	0.25	0.5	0.5	0.25	0.5	0.5	-1	0.5	1	1	1
WLC	1	1	1	1	1	0.5	0.5	0.25	0.5	0.5	0.25	0.5	0.5	0.25	0.5	0.5	-1	1	1	1
SPS	0.6	0.8	0.86	1	0.76	0.9	1	1	0.9	1	1	0.9	1	1	0.9	1	1	-1	0.29	0.28
SPN	0.67	0.84	0.88	1	0.8	0.92	1	1	0.92	1	1	0.92	1	1	0.92	1	1	0.29	-1	0.02
SPW	0.67	0.84	0.88	1	0.8	0.92	1	1	0.91	1	1	0.92	1	1	0.92	1	1	0.28	0.02	-1

شکل ۲۳-۴. ماتریس AIDM

در ماتریس شکل ۲۳-۴ سطر و یا ستون‌ها بر اساس ترتیب الگوریتم‌ها در جدول ۲-۴ مرتب شده‌اند و کد هر یک از الگوریتم‌ها جهت خوانایی بیشتر در کنار سطر و یا ستون متعلق به آن نوشته شده است. در این جدول درجه استقلال برای مقایسه هر الگوریتم با خودش به صورت قراردادی منفی یک در نظر گرفته شده است که همانند روش پیشنهادی اول باید ازتابع BPI حین اجرای الگوریتم برای محاسبه درجه استقلال آن بر اساس پارامترهای اساسی الگوریتم استفاده کنیم. قابل ذکر است که این ماتریس نسبت به قطر اصلی آن متقارن است از این روی در ذخیره‌سازی فقط و فقط بخش بالا و یا پایین قطر اصلی را در حافظه ذخیره می‌کنیم.

۴-۵. نتایج آزمایش‌ها

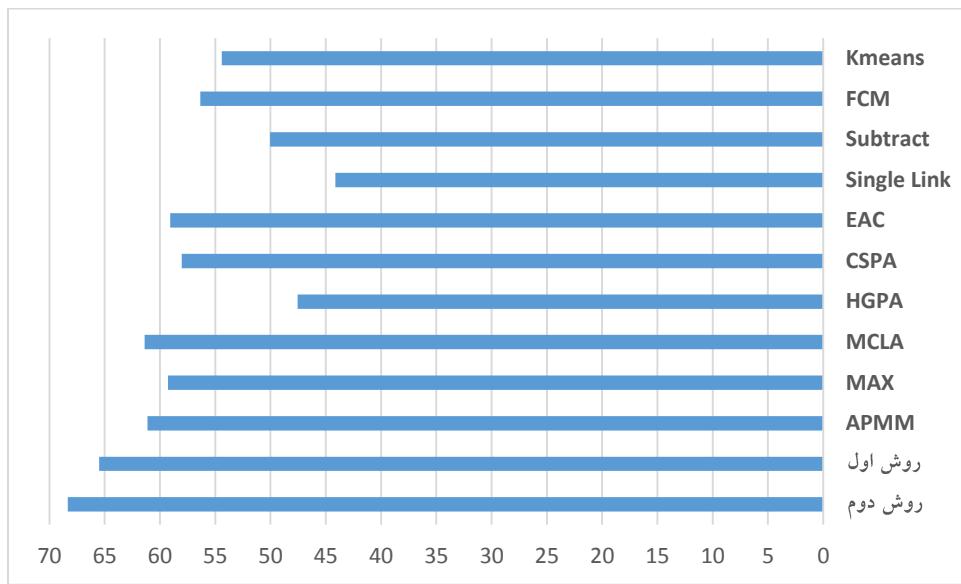
در این تحقیق همان طور که پیش‌تر اشاره شد از نرم‌افزار Matlab R2013a (8.1.0.604) جهت پیاده‌سازی الگوریتم‌ها و ارزیابی آن‌ها استفاده شده است. مقادیر آستانه استقلال، پراکندگی و عدم تمرکز در روش پیشنهادی اول و همچنین مقادیر آستانه پراکندگی و عدم تمرکز در روش پیشنهادی دوم طوری انتخاب شده است که با یک رایانه با مشخصات معلوم^{۱۸۷} اجرای الگوریتم‌ها به حداقل سه دقیقه زمان احتیاج داشته باشد. جهت اطمینان از صحت اجرای الگوریتم، هر یک از الگوریتم‌های پیشنهادی به طور مستقل ده بار اجرای شده‌اند و نتیجه‌ی نهایی میانگین نتایج ده بار اجرا فرض شده است. نتایج روش‌های پیشنهادی این تحقیق با نتایج چهار الگوریتم خوشه‌بندی پایه معروف که دارای تابع هدف متفاوت هستند و همچنین هفت الگوریتم خوشه‌بندی ترکیبی که دارای روش‌های ترکیب و انتخاب مختلف می‌باشند مقایسه شده‌اند. جدول ۴-۴ میزان دقیقت نتایج این الگوریتم‌های خوشه‌بندی را نسبت به کلاس‌های واقعی داده بر اساس درصد نشان می‌دهد.

جدول ۴-۴. دقیقت نتایج این الگوریتم‌های خوشه‌بندی را نسبت به کلاس‌های واقعی داده

	الگوریتم‌های خوشه‌بندی ترکیبی												روش پیشنهادی دوم	روش پیشنهادی اول	<i>iT</i>	<i>dT</i>	<i>cT</i>	دقیقت	<i>dT</i>	<i>cT</i>	درصد
	Kmeans	FCM	Sub tractive	Single Link	EAC	MAX	CSPA	HGPA	MCLA	APMM	<i>iT</i>	<i>dT</i>	<i>cT</i>								
Half Ring	75.75	78	86	75.75	77.17	78.48	74.5	50	74.5	80	0.2	0.06	3	87.2	0.21	6	97.8				
Iris	65.3	82.66	55.3	68	96	72.89	85.34	48.66	89.34	74.11	0.2	0.06	1	96	0.13	2	94.3				
Balance Scale	40.32	44	45.32	46.4	52	52.1	51.84	41.28	51.36	52.65	0.23	0.063	3	54.88	0.06	1	55.64				
Breast Cancer	93.7	94.43	65	65.15	95.02	75.72	80.97	50.37	96.05	96.04	0.18	0.02	1	96.92	0.05	2	97				
Bupa	54.49	50.1	57.97	57.68	55.18	56.17	56.23	50.72	55.36	55.07	0.21	0.04	3	57.42	0.07	3	57.83				
Galaxy	30.03	34.98	29.72	25.07	31.95	32.78	29.41	31.27	28.48	33.72	0.2	0.05	2	35.88	0.15	2	37.18				
Glass	42.05	47.19	36.44	36.44	45.93	44.17	38.78	41.12	51.4	47.19	0.19	0.06	3	51.82	0.06	3	50				
Ionosphere	69.51	67.8	77	64.38	70.48	64.48	67.8	58.4	71.22	70.94	0.3	0.1	3	70.52	0.12	3	72.14				
SA Heart	64.51	63.41	67.26	65.15	65.19	63.96	58.42	50.93	62.54	70.91	0.65	0.8	1	68.7	0.5	1	72.47				
Yeast	31.19	29.98	31.2	31.73	31.74	32.4	14	15.23	17.56	31.06	0.5	0.5	1	34.76	0.8	3	32.52				
Wine	65.73	71.34	67.23	37.64	70.56	69.17	67.41	62.36	70.22	64.6	0.2	0.05	3	71.34	0.1	1	94.55				
Pendigits	40.97	36.77	10.4	10.46	43.9	57.02	58.32	47.55	58.62	47.4	0.02	0.12	1	58.68	0.1	1	59.23				
Optdigit	47.23	38.33	47.72	10.28	48.12	76.11	75.21	64.77	77.15	77.1	0.01	0.1	1	77.16	0.1	1	77.97				
Statlog	40.89	49.91	23.8	23.8	43.96	54.23	54.23	52.94	55.71	54.88	0.01	0.1	1	55.77	0.08	1	57.86				

¹⁸⁷ CPU=Intel X9775 (4*3.2 GHz), RAM=16GB, OS=Windows Server 2012 RTM x64

همان طور که در جدول ۴-۴ مشاهده می‌شود در برخی از مجموعه داده‌ها همانند Bupa و یا Ionosphere، الگوریتم خوشبندی کاہشی نتیجه دقیق‌تری نسبت به سایر روش‌ها تولید کرده است ولی سایر روش‌های پایه قادر به تشخیص الگوهای دقیق در این داده‌ها نمی‌باشند. این مجموعه داده‌ها مثالی خوب از مسئله انبار کاه در نظریه خرد جمعی می‌باشد که در آن فقط یک الگوریتم، الگوهای صحیح را تشخیص داده و سایر الگوریتم‌های پایه الگوهایی با دقت بالا را تشخیص نداده‌اند (فقط تعداد اندکی از اعضای مجمع جواب صحیح را می‌دانند). همان‌گونه که در دقت نتایج این داده‌ها مشاهده می‌شود هیچ‌کدام از روش‌های ترکیبی نتوانسته‌اند الگوهای صحیح را با ترکیب نتایج اولیه شناسایی کنند. شکل ۲۴-۴ میانگین دقت روش‌های مختلف را بر روی داده‌های جدول ۴-۴ نشان می‌دهد.



شکل ۲۴-۴. میانگین دقت الگوریتم‌های خوشبندی

همان طور که در شکل ۲۴-۴ مشاهده می‌شود با اینکه در دو داده ذکر شده الگوریتم کاہشی بهتر عمل کرده است ولی در سایر نتایج دقت مناسبی نداشته است. این مسئله، مثال مناسبی برای نشان دادن وابستگی کارایی الگوریتم‌های خوشبندی پایه به جنبه‌های خاص از مجموعه داده و دلیلی محکم جهت به کار گیری روش‌های ترکیبی می‌باشد. علاوه بر آن، مطابق با شکل بالا با این که روش ترکیب کامل تمامی افزارهای نتایج اولیه را در تولید نتیجه نهایی استفاده کرده است ولی دقت کمتری را نسبت به سایر روش‌های ترکیبی می‌تبنی بر انتخاب داراست. این مسئله در داده‌های بزرگ بهتر خود را نشان می‌دهد به طوری که در سه مجموعه داده آخر جدول ۴-۴ تشخیص الگوی صحیح با ترکیب تمامی نتایج بسیار سخت حاصل می‌شود. از طرف دیگر چهارچوب غیرمت مرکز یکی از مهم‌ترین

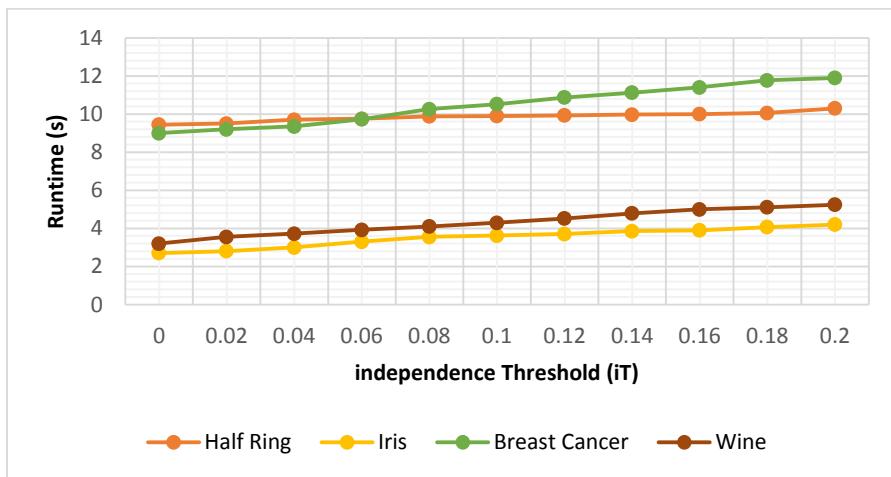
دلایل بهتر کار کردن الگوریتم های پیشنهادی در این نوع داده ها می باشد چون در سایر روش های مبتنی بر انتخاب تغییرات در مقادیر ارزیابی معیار پراکنده گی قبل و پس از انتخاب خوش بسیار بر روی دقت نتیجه نهایی تأثیر می گذارد. علاوه بر آن، می توان به تأثیرات سایر معیار های پیشنهادی در خوش بندی خردمند از جمله استقلال نیز اشاره کرد چون با اینکه روش های APMM و MAX نیز از معیاری مشابه با روش های پیشنهادی این تحقیق جهت ارزیابی پراکنده گی استفاده می کنند ولی نتایج روش های پیشنهادی مقاومه بهتر از آنها عمل کرده اند. قابل ذکر است که با اینکه روش دوم در مجموع تقریباً پنج درصد بهتر از روش اول نتایج را بهبود بخشید ولی در این روش دیگر نیاز به تعیین آستانه برای استقلال الگوریتم نیست و همچنین از نظر زمان اجرا بسیار بهتر از روش اول عمل می کند که در ادامه به تشریح آن خواهیم پرداخت. این تحقیق برای اطمینان از دقت نتایج حاصل در آزمایش های تجربی تمامی آزمایش ها را با استفاده از معیار اطلاعات متقابل نرمال شده نیز ارزیابی کرده است که نتایج آن در جدول ۴-۵ مشاهده می شود.

جدول ۴-۵. جدول مقایسه معیار اطلاعات متقابل نرمال شده (NMI) نتایج آزمایش

	الگوریتم های خوش بندی ترکیبی											روش دوم	روش اول
	Kmeans	FCM	Subtract	Single Link	EAC	MAX	CSPA	HGPA	MCLA	APMM			
Half Ring	0.26	0.33	0.51	0.06	0.32	0.56	0.32	0	0.34	0.32	0.68	0.74	
Iris	0.74	0.78	0.77	0.73	0.75	0.8	0.72	0.14	0.75	0.5	0.86	0.81	
Balance Scale	0.12	0.2	0.37	0.03	0.14	0.22	0.07	0.03	0.09	0.31	0.37	0.41	
Breast Cancer	0.74	0.69	0.73	0.01	0.69	0.72	0.35	0	0.74	0.7	0.74	0.74	
Bupa	0.0008	0.0045	0	0.0136	0.0018	0.002	0.01	0	0.0016	0.0009	0.0013	0.0017	
Galaxy	0.24	0.26	0.17	0.12	0.28	0.31	0.23	0.13	0.27	0.31	0.34	0.4	
Glass	0.36	0.24	0.07	0.11	0.41	0.3	0.25	0.31	0.27	0.4	0.45	0.46	
Ionosphere	0.12	0.08	0.07	0.02	0.11	0.12	0.1	0.02	0.13	0.13	0.15	0.12	
SA Heart	0.08	0.13	0.13	0	0.07	0.08	0.02	0	0.08	0.075	0.078	0.041	
Yeast	0.1	0.11	0	0.11	0.12	0.27	0.26	0.14	0.28	0.03	0.28	0.26	
Wine	0.75	0.55	0.72	0.05	0.69	0.79	0.77	0.43	0.8	0.64	0.83	0.84	
Pendigits	0.61	0.35	0	0.01	0.01	0.7	0.68	0	0.69	0.01	0.71	0.71	
Optdigit	0.6	0.39	0.45	0.02	0.36	0.76	0.68	0.42	0.73	0.79	0.76	0.79	
Statlog	0.52	0.38	0	0.01	0.01	0.54	0.47	0.42	0.56	0.54	0.56	0.53	

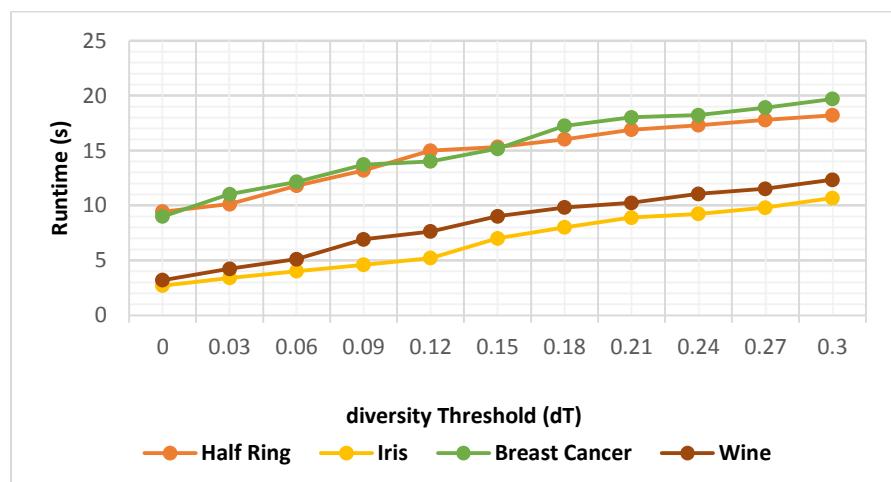
جدول ۴-۵ ارزیابی نتایج آزمایش های این تحقیق توسط معیار اطلاعات متقابل نرمال شده را نسبت به کلاس های واقعی داده نشان می دهد. همان گونه که مشاهده می شود در مجموعه داده Bupa روش ترکیبی CSPA و در مجموعه داده SA Heart الگوریتم خوش بندی کا هشی بیشترین مقدار NMI را داراست. مهم ترین دلیل این نتایج ریشه در روش حل مسئله این الگوریتم ها (تابع هدف و یا روش ترکیب) دارد به نحوی که در این داده ها روش های مذکور مقادیر NMI را بهتر از سایر روش ها بهبود می بخشنند ولی در بیشتر موارد الگوریتم های پیشنهادی این تحقیق دارای NMI مناسب تری نسبت به سایر الگوریتم ها می باشند. در ادامه به بررسی تأثیرات مقادیر آستانه بر روی دقت و زمان اجرای

الگوریتم‌های پیشنهادی خواهیم پرداخت. در شکل ۲۵-۴ رابطه میان مقدار آستانه استقلال و زمان اجرای الگوریتم در روش اول بررسی شده است. در این نمودار مقدار آستانه پراکنده‌گی صفر و عدم تمرکز یک فرض شده است تا تأثیر آنها بر روی نتایج حذف شود. در این شکل محور افقی مقادیر آستانه استقلال الگوریتم و محور عمودی زمان اجرا به ثانیه را نشان می‌دهد.



شکل ۲۵-۴. رابطه میان آستانه استقلال و زمان اجرای الگوریتم در روش پیشنهادی اول

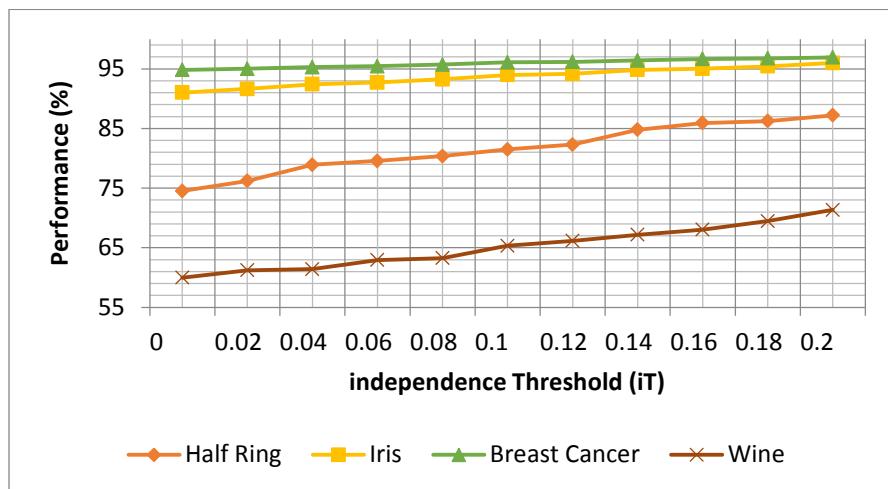
شکل ۲۶-۴ رابطه بین آستانه پراکنده‌گی در روش پیشنهادی اول و زمان اجرای برنامه را نشان می‌دهد. در این نمودار مقدار آستانه استقلال صفر و عدم تمرکز یک فرض شده است تا تأثیر آن بر روی نتایج حذف شود. در این شکل محور افقی آستانه پراکنده‌گی و محور عمودی زمان اجرا به ثانیه می‌باشد.



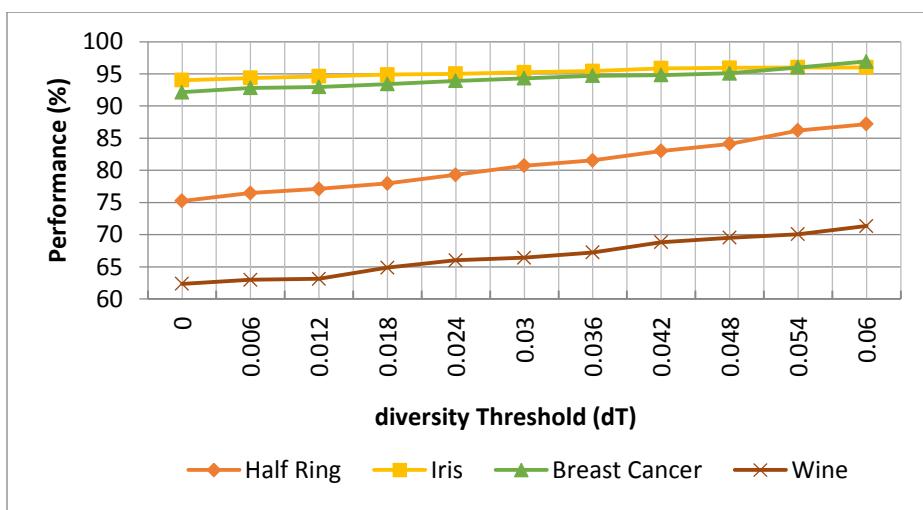
شکل ۲۶-۴. رابطه میان آستانه پراکنده‌گی و زمان اجرای الگوریتم در روش پیشنهادی اول

همان‌گونه که در شکل‌های ۲۵-۴ و ۲۶-۴ مشاهده می‌شود هر چه مقادیر آستانه استقلال و پراکنده‌گی در الگوریتم روش پیشنهادی اول بیشتر شود زمان اجرای الگوریتم نیز بیشتر خواهد شد. این موضوع

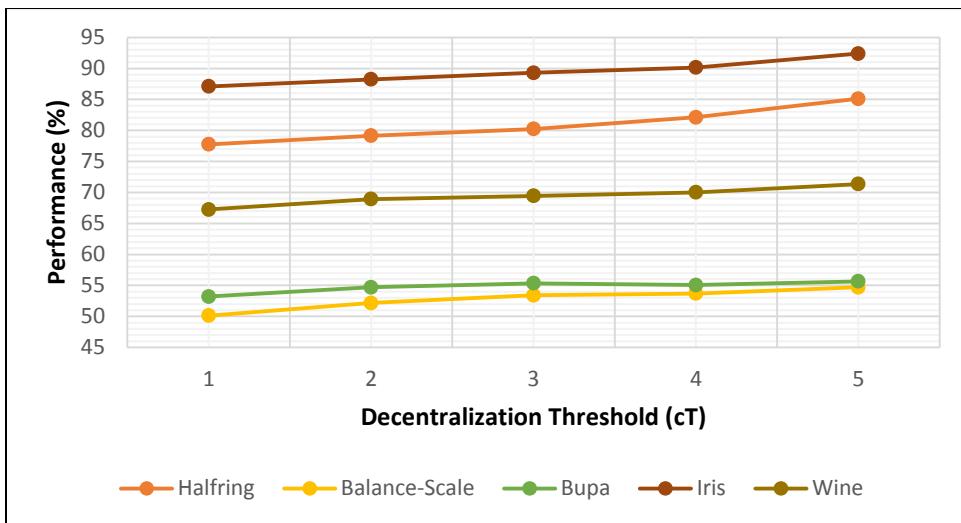
بدیهی است چون با بالا رفتن مقادیر آستانه باید تعداد بیشتری نتیجه اولیه تولید شود تا در آنها به دنبال الگوهای باکیفیت‌تر بگردیم. در ادامه به بررسی تأثیرات افزایش مقادیر آستانه بر روی دقت نتایج خواهیم پرداخت. شکل ۲۷-۴، ۲۸-۴ و ۲۹-۴ به ترتیب نشان‌دهنده رابطه دقت (کارایی) در روش پیشنهادی اول با مقادیر آستانه استقلال، پراکندگی و عدم تمرکز می‌باشد که در تمامی موارد جهت حذف تأثیرشان سایر مقادیر آستانه به ازای استقلال و پراکندگی صفر و برای عدم تمرکز یک در نظر گرفته شده‌اند. در این نمودارها محورهای افقی مقادیر آستانه و محورهای عمودی مقادیر دقت را به درصد نشان می‌دهد.



شکل ۲۷-۴. رابطه میان آستانه استقلال و دقت نتیجه نهایی در روش پیشنهادی اول

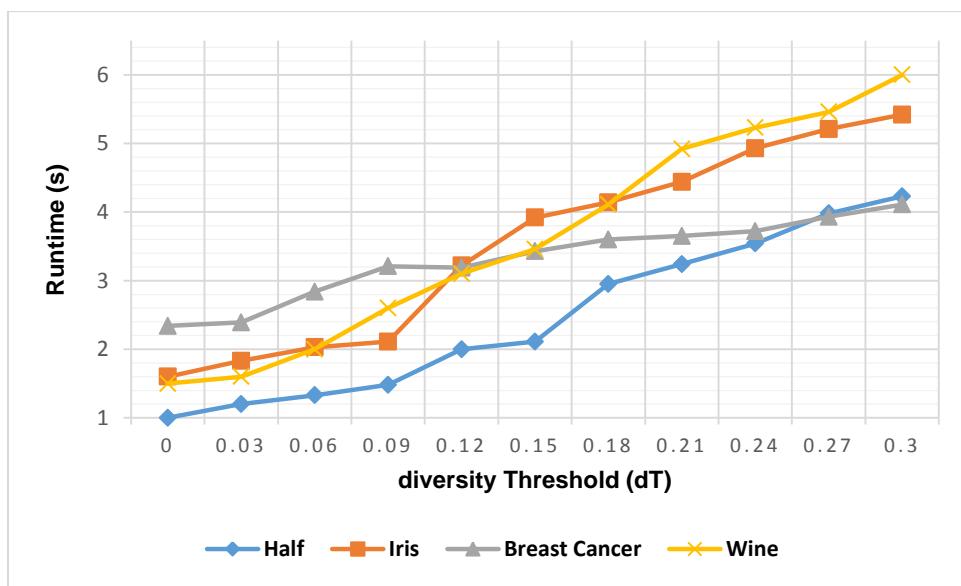


شکل ۲۸-۴. رابطه میان آستانه پراکندگی و دقت نتیجه نهایی در روش پیشنهادی اول



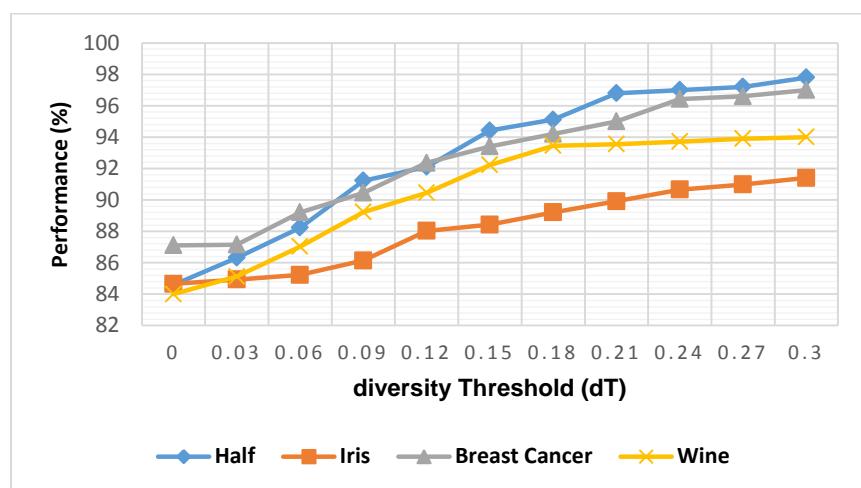
شکل ۲۹-۴. رابطه میان آستانه عدم تمرکز و دقت نتیجه نهایی در روش پیشنهادی اول

همانگونه که در شکل های بالا مشاهده می شود هر چه مقادیر آستانه در الگوریتم روش پیشنهادی اول بیشتر شود دقت نتایج الگوریتم نیز بیشتر خواهد شد. این موضوع بدیهی است چون با بالا رفتن مقادیر آستانه تعداد بیشتری الگوی باکیفیت تولید می شود که می تواند دقت مجمع خردمند را افزایش دهند. در ادامه به بررسی تأثیرات مقادیر آستانه بر روی زمان اجرا و دقت الگوریتم پیشنهادی دو می پردازیم. شکل ۳۰-۴ رابطه میان آستانه پراکندگی و زمان اجرای الگوریتم روش دوم را بررسی می کند. در این شکل محور افقی مقدار آستانه پراکندگی و محور عمودی زمان اجرا به ثانیه می باشد. در این آزمایش مقدار عدم تمرکز یک در نظر گرفته شده است تا تأثیر آن بر روی نتایج حذف شود.



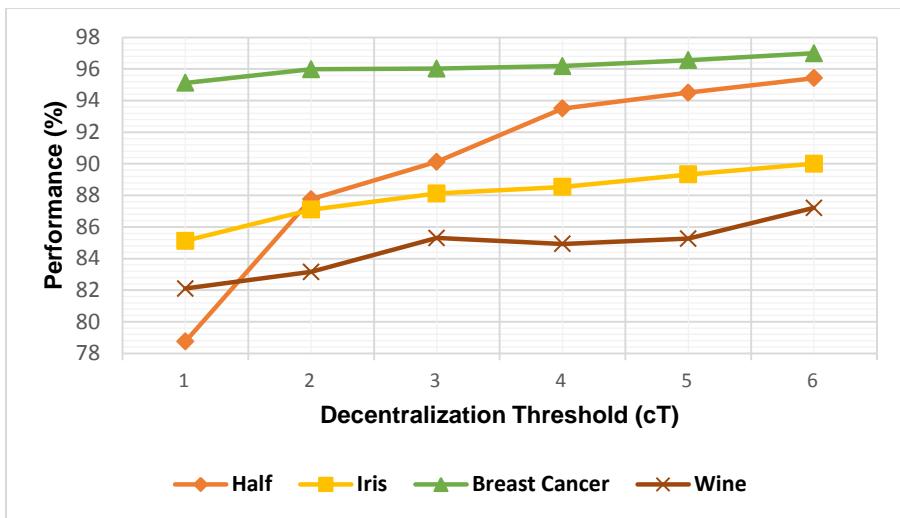
شکل ۳۰-۴. رابطه میان آستانه پراکندگی و زمان اجرای الگوریتم در روش پیشنهادی دوم

مطابق با مشاهدات شکل ۳۰-۴ هرچه مقادیر آستانه پراکندگی در روش پیشنهادی دوم بیشتر شود زمان اجرای الگوریتم افزایش پیدا می‌کند. همچنین این شکل نشان می‌دهد که در مقایسه با الگوریتم روش پیشنهادی اول، این الگوریتم دارای کمتر ولی حساسیت بیشتر نسبت به تغییرات آستانه پراکندگی می‌باشد که در ادامه به بررسی و مقایسه آن‌ها خواهیم پرداخت ولی قبل از آن به بررسی تأثیرات مقادیر آستانه بر روی دقت نتایج در روش پیشنهادی دوم می‌پردازیم. شکل ۳۱-۴ رابطه میان آستانه پراکندگی و دقت نتایج نهایی در روش دوم را بررسی می‌کند. در این شکل محور افقی مقدار آستانه پراکندگی و محور عمودی دقت الگوریتم در مقایسه با کلاس واقعی داده به درصد می‌باشد. در این آزمایش مقدار عدم تمرکز یک در نظر گرفته شده است تا تأثیر آن بر روی نتایج حذف شود.



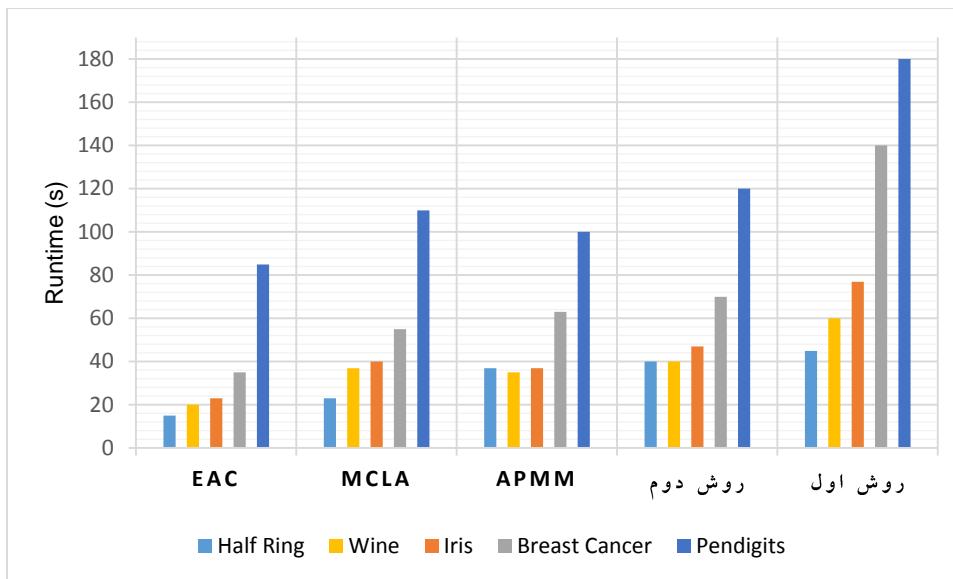
شکل ۳۱-۴. رابطه میان آستانه پراکندگی و دقت نتایج نهایی در روش پیشنهادی دوم

شکل ۳۲-۴ رابطه میان آستانه عدم تمرکز و دقت نتایج نهایی در روش پیشنهادی دوم را بررسی می‌کند در این شکل محور افقی مقدار آستانه عدم تمرکز و محور عمودی دقت الگوریتم در مقایسه با کلاس واقعی داده به درصد می‌باشد. در این آزمایش مقدار پراکندگی صفر در نظر گرفته شده است تا تأثیر آن بر روی نتایج حذف شود.



شکل ۳۲-۴. رابطه میان آستانه عدم تمرکز و دقت نتایج نهایی در روش پیشنهادی دوم

در شکل ۳۱-۴ و ۳۲-۴ هرچه مقادیر آستانه در روش پیشنهادی دوم بیشتر شود دقت نتایج نهایی افزایش پیدا می‌کند. مطابق با نمودارهای شکل‌های ۳۲-۴ الی ۲۵-۴ می‌توان به این نتیجه رسید که در الگوریتم‌های پیشنهادی این تحقیق می‌توان با بالا بردن مقادیر آستانه دقت نتایج را افزایش داد ولی این امر باعث می‌شود تا زمان اجرای الگوریتم نیز افزایش یابد. از این روی باید تعادلی بین کارایی (دقت) و زمان اجرای الگوریتم برقرار کرد. همان طور که پیش‌تر نیز اشاره شد در این تحقیق به صورت توافقی حداکثر زمان اجرای الگوریتم‌ها ثابت فرض شده‌اند و متناسب با آن به آستانه‌های الگوریتم را تنظیم می‌کنیم. نتیجه دیگری که از نمودارهای زمان اجرای دو الگوریتم می‌توان گرفت این است که زمان اجرای الگوریتم‌های پیشنهادی همانند سایر روش‌ها به اندازه و ابعاد داده نیز وابسته می‌باشد. در شکل ۳۳-۴ زمان اجرای الگوریتم‌های پیشنهادی این مقاله بر روی چند داده استاندارد را با روش ترکیب کامل (EAC) و دو روش خوشبندی ترکیبی دیگر که نتایج دقت آن‌ها در جدول ۴-۴ نسبت به سایر روش‌ها بهتر بوده است مقایسه شده است. در این نمودار محور عمودی زمان اجرای الگوریتم بر حسب ثانیه می‌باشد و محور افقی بر اساس الگوریتم‌های خوشبندی مذکور گروه‌بندی شده است که در هر یک از این گروه‌ها شامل داده‌های استاندارد ارزیابی شده با آن الگوریتم‌ها می‌باشند. در این آزمایش سعی شده از داده‌هایی با ابعاد کوچک، متوسط و بزرگ استفاده شود تا تأثیرات ابعاد داده بر روی نتایج آزمایش کم شود.



شکل ۴-۳۳. مقایسه زمان اجرای الگوریتم

همان‌گونه که در شکل ۴-۳۳ نشان داده شده است کمترین زمان اجرا برای روش خوش‌بندی ترکیبی کامل می‌باشد چون در این روش پس از تولید نتایج اولیه هیچ ارزیابی صورت نمی‌گیرد. همان‌گونه که می‌بینید روش پیشنهادی اول به خاطر ارزیابی استقلال به زمان اجرای بیشتری نیاز دارد ولی در روش پیشنهادی دوم با در نظر گرفتن معیار استقلال به عنوان یک ضریب برای پراکندگی در ترکیب نتایج اولیه می‌توان زمان اجرای الگوریتم را به مقدار قابل ملاحظه‌ای کاهش داد به نحوی که زمان اجرای الگوریتم دوم تقریباً دوست دارد از روش APMM بیشتر به طول می‌انجامد ولی دارای دقت بیشتری نسبت به آن می‌باشد. از این روی یکی دیگر از ویژگی‌های روش پیشنهادی دو نسبت به روش اول زمان اجرای کمتر (تقریباً سی درصد) می‌باشد ولی همیشه جهت پیاده‌سازی آن نیاز به مدل‌سازی گراف استقلال الگوریتم می‌باشد.

فصل پنجم

جمع‌بندی و

کارهای آینده

۵. جمع‌بندی و کارهای آینده

۱-۵. جمع‌بندی

در این تحقیق برای اولین بار ایده استفاده از نظریه خرد جمعی جهت انتخاب خوشه در خوشه‌بندی ترکیبی مبتنی بر انتخاب مطرح، و برای پیاده‌سازی آن دو الگوریتم پیشنهاد شده است که روش دوم روشی توسعه‌یافته بر اساس روش پیشنهادی اول می‌باشد. در این راستای ابتدا برای شرایط چهارگانه خرد جمعی (پراکنده‌گی، استقلال، عدم تمرکز و روش ترکیب مناسب) باز تعریفی متناسب با ادبیات خوشه‌بندی ترکیبی صورت گرفته و معیارهایی مرتبط با این تعاریف ارائه شده است. در معیار پراکنده‌گی مطرح شده در این تحقیق جهت رفع مشکل تقارن معیار کلاسیک اطلاعات متقابل نرمال شده (NMI) در ارزیابی پراکنده‌گی افزارها، یک معیار جدید بر اساس معیار APMM که برای ارزیابی پراکنده‌گی خوشه استفاده می‌شود با عنوان A3 ارائه شده است. از طرفی دیگر در این تحقیق برای اولین بار مفهوم درجه استقلال الگوریتم‌های خوشه‌بندی و تأثیرات آن بر روی نتایج نهایی مطرح شده که مطابق با این مفهوم در روش پیشنهادی اول، الگوریتم‌های غیر هم نام کاملاً مستقل و الگوریتم‌های هم نام بر اساس فرآیند آستانه‌گیری از درجه استقلال پارامترهای اساسی الگوریتم ارزیابی می‌شوند. در روش پیشنهادی دوم جهت ارزیابی درجه استقلال الگوریتم‌ها ابتدا زبان استقلال الگوریتم‌های خوشه‌بندی (CAIL) معرفی شده و سپس روشی مبتنی بر گراف کد الگوریتم جهت ارزیابی درجه استقلال الگوریتم‌های خوشه‌بندی با استفاده از کدهای تولیدشده به این زبان ارائه شده است. در این تحقیق برای خودکار سازی فرآیند ارزیابی درجه استقلال الگوریتم با استفاده از کد استقلال ابزاری با عنوان تحلیلگر کد استقلال الگوریتم (CAIL Code Analyzer) طراحی و معرفی می‌شود. همچنین در روش پیشنهادی دوم درجه استقلال الگوریتم به عنوان یک وزن برای پراکنده‌گی در تشکیل نتیجه نهایی مورد استفاده قرار گرفته است که برای رسیدن به این منظور روش انباشت مدارک وزن‌دار (WEAC) با توسعه روش انباشت مدارک معرفی شده است. در ادامه این تحقیق، چهارچوب غیرمت مرکز و فرآیند بازخورد جهت تولید نتایج اولیه خوشه‌بندی، ارزیابی و ترکیب آن‌ها معرفی و تأثیرات آن بر روی پراکنده‌گی افزارها قبل و بعد از انتخاب بیان می‌شود. جهت بررسی ادعاهای این تحقیق نتایج آزمایش‌های تجربی و تحلیل آن‌ها در فصل چهارم ارائه شده است که بهبود کلی نتایج نهایی در روش‌های پیشنهادی این تحقیق را نسبت به روش‌های پیشین نشان می‌دهد.

۲-۵. کارهای آینده

در این بخش لیستی از کارهایی که می‌توان در ادامه این تحقیق جهت توسعه یا به کارگیری روش‌های پیشنهادی مطرح شده انجام داد ذکر می‌شود.

- ۱- توسعه روش پیشنهادی مبتنی بر گراف جهت مدل‌سازی دقیق‌تر الگوریتم‌های خوشه‌بندی پایه و تسريع سرعت اجرای الگوریتم.
- ۲- خودکارسازی فرآیند تبدیل کد و یا شبیه کد الگوریتم‌ها به زبان استقلال الگوریتم‌های خوشه‌بندی.
- ۳- استفاده از الگوریتم‌های پیشنهادی این تحقیق در ارائه الگوریتم‌های خوشه‌بندی چندگانه.
- ۴- استفاده از داده و ویژگی‌های آن در فرآیند انتخاب خوشه در چهارچوب غیرمت مرکز.
- ۵- توسعه مفهوم استقلال الگوریتم‌های خوشه‌بندی و ارائه روش‌های دقیق‌تر برای مدل‌سازی آن.
- ۶- خودکارسازی تعیین مقادیر آستانه برای معیارهای پراکندگی و عدم تمرکز در روش‌های پیشنهادی
- ۷- خودکارسازی روش انتخاب تعداد خوشه در جواب نهایی بر اساس بهترین مقدار قابل تخصیص.

منابع و مأخذ

- [1] حسین علیزاده، خوشبندی ترکیبی مبتنی بر زیرمجموعه‌ای از نتایج اولیه، پایان‌نامه کارشناسی ارشد، دانشکده کامپیوتر دانشگاه علم و صنعت ایران، اسفند ۱۳۸۷
- [2] جواد عظیمی، بررسی پراکندگی در خوشبندی ترکیبی، پایان‌نامه کارشناسی ارشد، دانشکده کامپیوتر دانشگاه علم و صنعت ایران، خرداد ۱۳۸۶
- [3] Ayad H.G. and Kamel M.S., "Cumulative Voting Consensus Method for Partitions with a Variable Number of Clusters", IEEE Trans. On Pattern Analysis and Machine Intelligence, VOL. 30, NO. 1, 160-173, 2008.
- [4] Azimi J., Maani J. and Mozayyeni N., "Improved Clustering Ensembles", 11th International CSI Computer Conference (CSICC06), Tehran, Iran, pp. 24-26, January 2006.
- [5] Azimi J., Mohammadi M., Analoui M., "Clustering Ensembles Using Genetic Algorithm", in IEEE CAMPS, 2006.
- [6] Azimi J., Analoui M., "Improved Clustering Ensembles Using Maximal Similar Features and Non-Random K-means", The IASTED International Conference on Artificial Intelligence and Applications, Innsbruck, Austria, pp. 12 – 14, February 2007.
- [7] Azimi J., Fern X., "Adaptive Cluster Ensemble Selection", International Joint Conferences on Artificial Intelligence, 2009.
- [8] Alizadeh H., Minae i.M. and Parvin H., "A New Asymmetric Criterion for Cluster Validation", Springer-Verlag Berlin Heidelberg, pp.320-30, 2011.
- [9] Alizadeh H., Parvin H. and Parvin S., "A Framework for Cluster Ensemble Based on a Max Metric as Cluster Evaluator", International Journal of Computer Science (IAENG), pp.1-39, 2012.
- [10] Baker L. and Ellison D., "The wisdom of crowds — ensembles and modules in environmental modeling", Geoderma, 147, pp.1-7, 2008.
- [11] Chung, F. R. K., "Spectral Graph Theory", CBMS Regional Conference Series in Mathematics, vol. 92, 1997.
- [12] Chris Ding and Xiaofeng He., "K-means clustering via Principal Component Analysis", Proc. of Int'l Conf. Machine Learning (ICML'04), pp. 225-232. July 2004.
- [13] Dunn J. C., "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics, vol. 3, pp. 32-57, 1973.

- [14] **Davis L.**, "Handbook of Genetic Algorithms", Van Nostrand Reinhold, New York, ISBN-10: 0442001738, ISBN-13: 978-0442001735, 1991.
- [15] **Driessche R.V. and D. Roose**, "An improved spectral bisection algorithm and its application to dynamic load balancing, parallel Computing, vol. 21, 1995.
- [16] **Dempster A. P., Laird N. M. and Rubin D. B.**, "Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of the Royal Statistical Society B, vol. 39, pp. 1-22, 1997.
- [17] **Dudoit S. and Fridly J.**, "Bagging to improve the accuracy of a clustering procedure", Bioinformatics 19, pp.1090–1099, 2013.
- [18] **Fred A. L. N. and Jain A. K.**, "Data Clustering Using Evidence Accumulation", Proc. of the 16th Intl. Conf. on Pattern Recognition, ICPR02, Quebec City, pp. 276 – 280, 2002.
- [19] **Fred A. L. N. and Jain A. K.**, "Combining Multiple Clusterings Using Evidence Accumulation. IEEE Trans", Pattern Analysis and Machine Intelligence, vol. 27(6), pp. 835–850, 2005.
- [20] **Fred A. L. N. and Jain A. K.**, "Learning Pairwise Similarity for Data Clustering", Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06) pp. 7695-2521, 2006.
- [21] **Fred A. L. N. and Lourenco A.**, "Cluster Ensemble Methods: from Single Clusterings to Combined Solutions", Studies in Computational Intelligence (SCI), vol. 126, pp. 3–30, 2008.
- [22] **Fern X. Z. and Brodley C. E.**, "Random projection for high dimensional data clustering: A cluster ensemble approach", In Proceedings of the Twentieth International Conference on Machine Learning, pp. 186–193, 2003.
- [23] **Fern X. and Lin W.**, "Cluster Ensemble Selection", SIAM International Conference on Data Mining (SDM08), 2008.
- [24] **Goldberg D. E.**, "Genetic Algorithms in Search", Optimization and Machine Learning, Addison-Wesley, New York, 1989.
- [25] **Ghahramani Z. and Jordan M.**, "Supervised learning from incomplete data via an EM approach", In Proceedings of Advances in Neural Information Processing Systems (NIPS 6), pp. 120-127, 1993.
- [26] **Gose E., Johnsonbaugh R., Jost S.**, "Pattern Recognition and Image Analysis", Prentice-Hall PTR, Upper Saddle River, NJ 07458, 1996.

- [27] **Grofman B. and Owen G.**, "Information Pooling and Group Decision Making", Proceedings of the Second University of California, Irvine Conference on Political Economy. JAI Press, Inc., Greenwich, Connecticut, 1996.
- [28] **Hagen, L. and Kahng A. B.**, "New spectral methods for ratio cut partitioning and clustering", Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, vol. 11(9), pp. 1074-1085, 1992.
- [29] **Hastie T., Tibshirani R. and Friedman J.**, "The Elements of Statistical Learning: Data Mining, Inference and Prediction", Springer, ISBN: 5-95284-387-0, 2003.
- [30] **Hong Q. Y. and Kwong S.**, "A genetic classification method for speaker recognition", Engineering Applications of Artificial Intelligence, vol. 18, pp. 13–19, 2005.
- [31] **Higham, D. J., Kalna G. and Kibble M.**, "Spectral clustering and its use in bioinformatics", Journal of Computational and Applied Mathematics, vol. 204(1), pp. 25-37, 2007.
- [32] **Hadzikadic M. and Sun M.**, "Wisdom of Crowds in the Prisoner's Dilemma Context", presented at Advances in Machine Learning II, pp.101-118, 2010.
- [33] **Jain A. K., and Dubes R. C.**, "Algorithms for Clustering Data. Englewood Cliffs", NJ: Prentice-Hall, 1998.
- [34] **Jain A., Murty M. N., and Flynn P.**, "Data clustering: A review", ACM Computing Surveys, vol. 31(3), pp. 264–323, 1999.
- [35] **Jianbo S. and J. Malik**, "Normalized cuts and image segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 22(8), pp. 888-905, 2000.
- [36] **Johnson S.**, "Emergence: the connected lives of ants, brains, cities and software", Scribner, ISBN 0-684-86876-8, 2002.
- [37] **Jain A. K., Topchy A., Law M., Buhmann J.**, "Landscape of Clustering Algorithms", In Proceedings of the 17th International Conference on Pattern Recognition, Cambridge UK, pp. 23-26, August 2004.
- [38] **Kernighan B. and Lin S.**, "An efficient heuristic procedure for partitioning graphs", Bell Systems Technical Journal, vol. 49, pp. 291-307, 1970.
- [39] **Kortsarz G and Peleg D.**, "On choosing a dense subgraph", In Proceedings of the 3th Annual IEEE Symposium on Foundations of Computer Science, pp. 692–701, 1993.
- [40] **Karypis G. and Kumar V.**, "A fast and high quality multilevel scheme for partitioning irregular graphs", SIAM Journal on Scientific Computing, vol. 20(1) pp. 359-392, 1998.

- [41] **Karypis G., Han E. H. and V. Kumar**, "Chameleon: Hierarchical clustering using dynamic modeling", IEEE Computer, pp. 32(8), pp. 68-75, August 1999.
- [42] **Kuncheva L. I. and Whitaker C. J.**, "Measures of diversity in classifier ensembles", Machine Learning, 2003.
- [43] **Langley P., Iba W. and Thompson K.**, "An analysis of Bayesian classifiers", In Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA, AAAI Press, pp. 399–406, 1992.
- [44] **Mackey C.**, "Extraordinary Popular Delusions and the Madness of Crowds", Natural, ISBN: 978-1897597323, 1841.
- [45] **McLachlan G. and Krishnan T.**, "The EM Algorithm and Extensions", Wiley, New York, 1997.
- [46] **Mitchell T. M.**, "Machine Learning", McGraw-Hill Companies, Inc., ISBN 0-07-042807-7, 1997.
- [47] **Monti S., Tamayo P., Mesirov J. and Golub T.**, "Consensus clustering: a resampling based method for class discovery and visualization of gene expression microarray data", Machine Learning, vol. 52, pp.91–118, 2003.
- [48] **Minaei-Bidgoli B., Punch W. F.**, "Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System", GECCO, pp. 2252-2263, 2003.
- [49] **Michael J.**, "Explaining the Wisdom of Crowds", Mauboussin on Strategy, Legg Mason Capital Management, 2007.
- [50] **Mok P. Y., Huang H. Q., Kwok Y. L. and Au J. S.**, "A robust adaptive clustering analysis method for automatic identification of clusters", Pattern Recognition, vol. 45, pp. 3017-3033, 2012.
- [51] **Ng A. Y., Jordan M. and Weiss Y.**, "On spectral clustering: analysis and an algorithm", Advances in Neural Information Processing Systems, vol. 14, 2002.
- [52] **Pothen, A.**, "Graph partitioning algorithms with applications to scientific computing", Parallel Numerical Algorithms. Kluwer, 1997.
- [53] **Page S. E.**, "The difference: How the power of diversity creates better groups, firms, schools, and societies", Princeton University Press, 2007.
- [54] **Strehl A. and Ghosh J.**, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions". Journal on Machine Learning Research, 3, pp. 583-617, 2002.
- [55] **Surowiecki J.**, "The Wisdom of Crowds", Cover of mass market edition by Anchor, ISBN: 978-0385503860, 2004.

- [56] **Topchy A., Jain A. K. and Punch W. F.**, "Combining Multiple Weak Clusterings", Proc. 3d IEEE Intl. Conf. on Data Mining, pp. 331-338, 2003.
- [57] **Topchy A., Minaei-Bidgoli B., Jain A. K. and Punch W. F.**, "Adaptive Clustering Ensembles", In Proc. Intl. Conf on Pattern Recognition, ICPR, Cambridge, UK, 2004.
- [58] **Topchy A., Jain A. K. and Punch W.**, "A Mixture Model for Clustering Ensembles", In Proceedings of the SIAM International Conference on Data Mining, Key: citeulike:3382369, 2004.
- [59] **Tritchler D., Fallah S. and Beyene J.**, "A spectral clustering method for microarray data", Computational Statistics & Data Analysis, vol. 49(1), pp. 63-76, 2005.
- [60] **Ward P. J.**, "Some Developments on the Affected-Pedigree-Member Method of Linkage Analysis", American journal of human genetics, Vol. 52, pp. 1200-1215, 1993.
- [61] **Xiang, T. and Gong S.**, "Spectral clustering with eigenvector selection", Pattern Recognition, vol. 41(3), pp.1012-1029, 2008.
- [62] **Zadeh L. A.**, "Fuzzy sets", Information and Control, vol. 8(3): 338–353. Doi: 10.1016/S0019-9958(65)90241-X. ISSN 0019-9958, 1965.
- [63] **Zhang T.**, "A unifying framework for spectral analysis based dimensionality reduction", Neural Networks, IJCNN (IEEE World Congress on Computational Intelligence), pp. 1670-1677, 2008.
- [64] **Zhao, F., Jiao L., Liu H., Gao X. and Gong M.**, "Spectral clustering with eigenvector selection based on entropy ranking", Neurocomputing, vol. 73(10-12), pp. 1704-1717, 2010.
- [65] **Alpert C. J. and Kahng A. B.**, "Recent directions in net list partitioning: A survey", Integration: The VLSI Journal, vol. 19, pp. 1-18, 1995.
- [66] **Bezdek J. C.**, "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York, 1981.
- [67] **Alizadeh H., Minaei-Bidgoli B. and Parvin H.**, "Cluster Ensemble Selection Based on a New Cluster Stability Measure", Intelligent Data Analysis, IOS Press, ISI Expanded, in press, will be appeared in Vol 18(3), 2014.
- [68] **Tan P. N., Steinbach M. & Kumar V.**, "Introduction to Data Mining", Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, ISBN:0321321367, 2005.
- [69] **Hendrickson B. and Leland R.**, "An improved spectral graph partitioning algorithm for mapping parallel computations ", SIAM Journal on Scientific Computing, 1995.

- [70] **Pothen A.**, "Graph Partitioning Algorithms with Applications to Scientific Computing", Parallel Numerical Algorithms, ICASE/LaRC Interdisciplinary Series in Science and Engineering, Vol. 4, pp. 323-368, 1997.
- [71] **Cover T. M. and Thomas J. A.**, "Elements of Information Theory", John Wiley & Sons, Inc., Print ISBN 0-471-06259-6, Online ISBN 0-471-20061-1, 1991.
- [72] **Albalate A. and Minker W.**, "Semi-Supervised and Unervised Machine Learning: Novel Strategies", John Wiley & Sons, Inc., Print ISBN 978-1-84821-203-9, 2013.
- [73] **Meila M.**, "Comparing Clusterings by the Variation of Information", Proceedings of COLT, pp 173-187, 2003.
- [74] **Jain A. K.**, "Data clustering: 50 years beyond K-means", Pattern Recognition Letters, pp. 1-16, 2009.
- [75] **Ammann P. and Offutt J.**, "Introduction to Software Testing", Cambridge press, ISBN: 9780521880381, 2008.
- [76] **Newman, C.B.D.J., Hettich, S. and Merz, C.**, "UCI repository of machine learning databases", [Online] Available at: <http://www.ics.uci.edu/~mlearn/MLSummary.html>, 1998.
- [77] **Zhang T., Szlam A. and Lerman G.**, "Median K-flats for hybrid linear modeling with many outliers", 12th IEEE International Conference on Computer Vision, 2009.
- [78] **Ledolter J.**, "Data mining business analytics with R", published by John Wiley & Sons, Inc., Hoboken, New Jersey, ISBN 978-1-118-44714-7, 2013.
- [79] **Sidhu R. S., Khullar S., Sandhu P. S., Bedi R. P. S. and Kaur K.**, "A Subtractive Clustering Based Approach for Early Prediction of Fault Proneness in Software Modules", World Academy of Science, Engineering and Technology, Vol. 43, 2010.
- [80] **Ulrike V. L.**, "A tutorial on spectral clustering", Statistics and computing, Springer, Vol. 17, Issue 4, pp. 395-416, 2007.
- [81] **Jordan F. and Bach F.**, "Learning spectral clustering", Advances in neural information processing systems, Vol. 16, 2003.
- [82] **Dhillon I. S., Guan Y. and Kulis B.**, "Kernel k-means: spectral clustering and normalized cuts", Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, publish by ACM, pp. 551-556, 2004.
- [83] **Verbeek J. J., Vlassis N. and Kroese B.**, "Efficient Greedy Learning of Gaussian Mixture Models", Neural Computation, Vol. 15(2), pp. 469-485, 2003.
- [84] **Hadjitodorov S. T., Kuncheva L. I. and Todorova L. P.**, "Moderate diversity for better cluster ensembles", Information Fusion, Vol. 7(3), pp. 264–275, 2006.

- [85] **Limin L. and Xiaoping F.**, "A New Selective Clustering Ensemble Algorithm", 9th IEEE International Conference on e-Business Engineering, 2012.
- [86] **Jia J., Xiao X. and Liu B.**, "Similarity-based Spectral Clustering Ensemble Selection", 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2012.
- [87] **Naldi M. C., Carvalho A. C. P. L. F. and Campello R. J. G. B.**, "Cluster ensemble selection based on relative validity indexes", Data Mining & Knowledge Discovery, Vol. 27, pp. 259–289 DOI 10.1007/s10618-012-0290-x.
- [88] **Yi S. K. M., Steyvers M. and Lee M. D.**, "Wisdom of the Crowds in Minimum Spanning Tree Problems", In: 32nd Annual Conference of the Cognitive Science Society, Austin, TX, USA: Cognitive Science Society, pp. 31840-31845, 10-13 August 2010.
- [89] **Welinder P., Branson S., Belongie S. and Perona P.**, "The Multidimensional Wisdom of Crowds", In: 24th Conference on Neural Information Processing Systems (NIPS), Vancouver, British Columbia, Canada: Curran Associates, Inc. pp. 1-9, 6-9 December 2010.
- [90] **Williams, D. P.**, "Underwater Mine Classification with Imperfect Labels", In: 20th International Conference on Pattern Recognition, Istanbul, Turkey. pp. 4157-4161, 2010.
- [91] **Miller B., Hemmer P., Steyvers M. and Lee M. D.**, "The Wisdom of Crowds in Rank Ordering Problems", In: 9th International Conference on Cognitive Modeling (ICCM'09), Manchester, UK: IEEE. pp. 86-91, 24-26 July 2009.
- [92] **Steyvers M., Lee M., Miller B. and Hemmer P.**, "The Wisdom of Crowds in the Recollection of Order Information", Advances in Neural Information Processing Systems, Vol. 22, pp.1785-93, 2009.

Abstract

One of the main tasks of data mining, clustering is used to group *non-labeled* data to find meaningful patterns. Generally, different models provide predictions with different accuracy rates. Thus, it would be more efficient to develop a number of models using different data subsets, or utilizing differing conditions within the modeling methodology of choice. However, selecting the best model is not necessarily the ideal choice because potentially valuable information may be wasted by discarding the results of less-successful models. This leads to the concept of combining, where outputs (individual predictions) of several models are pooled to make a better decision (collective prediction). Research in the Clustering Combination field has shown that these pooled outputs have more strength, novelty, stability, and flexibility than the results provided by individual algorithms.

Nevertheless, in the social science arena, there is a corresponding research field known as the Wisdom of Crowds, after the book by the same name written by Surowiecki in 2004, simply claiming that the Wisdom of Crowds (WOC) is the phenomenon whereby the decisions made by aggregating the information of groups usually have better results than those made by any single group members. Surowiecki suggested a clear structure for building a wise crowd. Supported by many examples from businesses, economies, societies, and nations, he argued that a wise crowd must satisfy four conditions, namely: diversity, independence, decentralization, and an aggregation mechanism.

This research studies the previous background and related work of cluster ensemble. Furthermore there is a review over the literature of the wisdom of crowds. The purpose of this study is to suggest ways of mapping and using this theory in selecting suitable clusters in the cluster ensemble. Thus, in this study two methods are proposed for combining the two techniques. Firstly, by incorporating the definitions used in the wisdom of crowds, in other words the four conditions of a wised crowd are redefined to be used in the cluster ensemble selection. Then, by using these definitions, the first method will be proposed in which thresholding is used for generating the final result. In this method the primary clustering algorithms with deferent types are considered independently, for which thresholding is needed. In the second method, the two parts of the first approach have been improved. In order to model and evaluate the independence of clustering algorithms, a technique based on algorithm graph code is presented. The degree of obtained independence level from this approach is used as a weight to evaluate diversity in generating the final result. To clarify our claim in this research, the results from the above approaches are compared with basic clustering methods, cluster ensemble methods, and cluster ensemble selection methods, using primarily standard UCI repository data sets. In conclusion section, all proposed methods for future work are also mentioned.

Keywords Cluster ensemble, Wisdom of crowd, Independency of algorithms, Diversity of results, Decentralization of framework



Mazandaran University of
Science and Technology

Cluster ensemble selection based on the wisdom of crowds

A Thesis Submitted in Partial Fulfillment of the Requirement for the Master's degree

Department Information Technology

By Muhammad Yousefnezhad

Supervisor Dr. Behrouz Minaei-Bidgoli

Advisor Mr. Hosein Alizadeh

September 2013

