روش اکتشافی جدید جهت خوشه‌بندی ابتدایی الگوریتم k-means

سید عسگری قاسمپوری[۱]، احمد برآنی[۲]، بهروز ترک لادانی[۳]

[۱] مربی، گروه کامپیوتر، دانشگاه آزاد اسلامی واحد قائمشهر، قائمشهر
a.ghasempouri@qaemshahriau.ac.ir

[۲] دانشیار، دانشکده کامپیوتر، دانشگاه اصفهان، اصفهان
ahmadb@eng.ui.ac.ir

[۳] دانشیار، دانشکده کامپیوتر، دانشگاه اصفهان، اصفهان
ladani@eng.ui.ac.ir

چکیده

با رشد روز افزون داده‌ها لزوم استخراج الگوهای مفید از آن‌ها بیشتر حس می‌گردد. یکی از روش‌های کشف دانش که بسیار مورد استفاده قرار می‌گیرد خوشه‌بندی می‌باشد. خوشه‌بندی به روش‌های مختلفی از جمله سلسله مراتبی و تکراری انجام می‌گیرد. در الگوریتم‌های خوشه‌بندی تکراری یکی از مهم‌ترین مراحل، انتخاب خوشه‌های اولیه است زیرا تاثیر مستقیم بر خوشه‌های نهایی دارد. از آنجایی‌که هر خوشه شامل نقاطی نزدیک به هم و دور از نقاط خوشه‌های دیگر است، انتخاب خوشه‌های اولیه اهمیت زیادی دارد. در این مقاله روشی اکتشافی و تکراری افزایشی برای تعیین خوشه‌های اولیه در الگوریتم k-means طراحی نمودیم. در هر مرحله دو عنصر جدید را برای خوشه‌ها انتخاب می‌کنیم. در ابتدا با یک خوشه که شامل یک عنصر می‌باشد کار خودر را آغاز کرده و در هر مرحله فاصله‌ی سایر عناصر با مرکز خوشه‌های تعیین شده را محاسبه می‌کنیم. این فاصله معیاری جهت تعیین عناصر خوشه‌های بعدی است. در این مقاله الگوریتم خود را بر روی چند مجموعه داده‌ی مختلف در اندازه‌های متفاوت اعمال کردیم. نتایج به‌دست آمده نشان می‌دهد روش ارائه شده باعث بهبود عملکرد الگوریتم k-means نسبت حالتی است که از خوشه‌های اولیه‌ی تصادفی استفاده شده است.

کلمات کلیدی

الگوریتم k-means، خوشه‌های اولیه، خوشه‌بندی، کشف دانش

۱- مقدمه

یکی از ابزارهای مهم داده‌کاوی و تحلیل داده‌های آماری خوشه‌بندی است که دارای کاربردهای فراوانی می‌باشد. هدف خوشه‌بندی این است که داده‌ها را به خوشه‌هایی تقسیم کنیم تا داده‌های درون یک خوشه دارای بیشترین شباهت و داده‌های خوشه‌های مختلف دارای کمترین شباهت باشند. روش‌های مختلفی برای خوشه‌بندی موجود است که به دو دسته‌ی سلسله‌مراتبی و غیر سلسله‌مراتبی تقسیم می‌گردند.

روش‌های سلسله‌مراتبی داده‌ها را به خوشه‌های سلسله‌مراتبی تودرتو تقسیم می‌کنند که با درخت دندوگرام قابل نمایش است. روش‌های غیرسلسله‌مراتبی داده‌ها را بر حسب شباهتشان به‌هم درون خوشه‌های مختلف قرار می‌دهند به‌طوری‌که داده‌های متشابه درون یک خوشه و داده‌های نامتشابه درون خوشه‌های مجزا قرار گیرند. این خوشه‌ها ممکن است دارای داده‌های مشترک نیز باشند.

الگوریتم k-means [۱] یکی از معروف‌ترین و سریع‌ترین روش‌های خوشه‌بندی می‌باشد. این الگوریتم غیر سلسله‌مراتبی است. سادگی k-means باعث شده است که در زمینه‌های مختلف مورد استفاده قرار گیرد. روش کار

به‌این صورت است که داده‌ها به k خوشه‌ی مجزا تقسیم می‌گردند. سپس در هر مرحله سعی می‌گردد با جابه‌جا کردن داده‌ها از خوشه‌ای به خوشه‌ی دیگر میانگین فاصله‌ی عناصر یک خوشه تا مرکز آن خوشه کمینه گردد. از آنجایی‌که k-means می‌تواند داده‌های حجیم را به صورتی کارا خوشه‌بندی نماید الگوریتمی بسیار پرکاربرد می‌باشد. اما این الگوریتم به انتخاب اولیه‌ی خوشه‌ها بسیار حساس است، به‌طوری‌که در صورت انتخاب تصادفی خوشه‌های اولیه، خوشه‌بندی حاصل از این الگوریتم همواره یکسان نبوده و دارای کیفیت‌های متفاوتی می‌باشد. در بعضی اوقات به‌دلیل انتخاب بد خوشه‌های اولیه، الگوریتم به مقدار بهینه‌ی محلی ختم می‌گردد. همچنین ممکن است به‌دلیل فاصله‌ی دور نقاط به بعضی مراکز خوشه، آن خوشه‌ها خالی بماند. بنابراین مهم‌ترین فاز این الگوریتم انتخاب خوشه‌های اولیه‌ی مناسب می‌باشد.

روش‌های مختلفی برای خوشه‌بندی اولیه در k-means پیشنهاد شده است. در [۲] روشی بازگشتی برای آغازین‌دهی خوشه‌های اولیه ارائه شده است. در [۳] الگوریتم k-mean چندین بار با خوشه‌های اولیه‌ی تصادفی اجرا می‌شود و میانگین مقادیر خوشه‌ها برابر مراکز خوشه‌ی نهایی خواهد شد.

در [۴] الگوریتم پالایشی ارائه شد که ابتدا با نمونه‌گیری تصادفی مجموعه‌ای از داده را تهیه می‌نمود. سپس با در نظر گرفتن k عدد از این مجموعه داده‌ها به عنوان k مرکز ثقل[۱]، داده‌های این مجموعه را خوشه‌بندی می‌کرد. این عمل با انتخاب مرکز ثقل‌های مختلف از مجموعه نمونه‌برداری شده تکرار می‌شد و مرکز ثقلی که خوشه‌های نهایی با حداقل خطای خوشه‌بندی را داشت به عنوان مرکز ثقل‌های کاندید انتخاب می‌گردید.

مقاله‌ی [۵] روشی مبتنی برتکرار را پیشنهاد داد که در آن به‌صورت پویا هربار یک مرکز خوشه‌ی جدید اضافه می‌گردد. یافتن این مراکز خوشه به وسیله‌ی جستجوی سراسری که شامل N اجرای الگوریتم k-means با مکان‌های آغازین مناسب است انجام می‌گیرد.

در [۶] روشی موسوم به الگوریتم آغازین‌دهی مرکز خوشه (CCIA) ارائه شد که مشکل آغازین‌دهی خوشه‌ها را حل می‌نمود. این روش بر اساس مشاهده‌ی دو معیار، الگوهای داده‌ای مشابه را تعیین می‌نماید. در این روش ابتدا میانگین و انحراف از معیار صفات داده‌ها محاسبه می‌گردد سپس داده‌ها به وسیله‌ی نمودار نرمال به بخش‌های خاصی تقسیم می‌شوند. در گام آخر به‌وسیله‌ی الگوریتم k-means و چگالش داده‌ی مبتنی بر تراکم شباهت الگوهای داده‌ای جهت تعیین خوشه‌های اولیه مشاهده می‌گردد. نتایج آزمایش‌ها نشان داده که این روش در بسیاری از روش‌های خوشه‌بندی کارایی مناسبی دارد.

مقاله‌ی [۷] الگوریتمی جهت تعیین مراکز خوشه‌ی اولیه برای k-means ارائه داده است. این روش با برش صفحه‌ی شامل مجموعه‌داده‌ها آن‌ها را به دو مجموعه‌ی مجزا به نام سلول تقسیم می‌کند. این صفحه عمود بر محور داده‌ای است که بالاترین واریانس را دارد و به این هدف تقسیم را انجام می‌دهد که همزمان مجموع میانگین مربعات خطای دو سلول را کمینه نموده و فاصله‌ی دو سلول را بیشینه نماید.

در ادامه، مقاله به شکل زیر سازمان یافته است. در بخش ۲ الگوریتم پیشنهادی برای محاسبه‌ی خوشه‌های آغازین روش k-means شرح داده شده است. در بخش ۳ معیار ارزیابی مورد استفاده برای تعیین کیفیت خوشه‌بندی بیان شده است. در بخش ۴ نتایج آزمایشات برروی یک‌سری مجموعه داده جمع‌آوری شده که در آن الگوریتم پیشنهادی به مجموعه داده‌های Iris، Wine و Abalone اعمال گشته است. در نهایت بخش ۵ شامل نتیجه‌گیری می‌باشد.

۲- الگوریتم پیشنهادی

در این بخش الگوریتم پیشنهادی جهت به‌دست آوردن خوشه‌های آغازین برای روش خوشه‌بندی k-means را شرح می‌دهیم. روش ما که یک روش تکرارای افزایشی است، در هر مرحله دو عنصر جدید را برای خوشه‌ها بر می‌گزینند. این روش در ابتدا با یک خوشه که شامل یک عنصر می‌باشد کار خودر را آغاز کرده و در هر مرحله فاصله‌ی سایر عناصر با مرکز خوشه‌های تعیین شده را محاسبه می‌کند. این فاصله معیاری جهت تعیین عناصر خوشه‌های بعدی است. روند اجرای الگوریتم در ادامه شرح داده شده است.

اولین مرحله‌ی الگوریتم تعیین عنصر آغازین است. در روش ما اولین عنصر مجموعه داده به عنوان عنصر آغازین در نظر گرفته می‌شود. عنصر آغازین به عنوان تنها عضو خوشه‌ی اول تعیین می‌گردد و از مجموعه داده اولیه حذف می‌شود تا در محاسبات بعدی درنظر گرفته نشود. پس از تخصیص هر عنصر به یک خوشه مرکز ثقل آن خوشه به‌روز می‌گردد. مرکز ثقل یک خوشه میانگین داده‌های متعلق به آن خوشه است. این کار با استفاده از فرمول (۱) انجام می‌شود. باید توجه داشت که این فرمول به تک‌تک صفات مرکز ثقل و داده‌ی جدید اعمال می‌گردد.

$$mean_{new}(c) = \frac{mean_{old}(c) * size_{old}(c) + data_{new}}{size_{old}(c) + 1} \qquad (۱)$$

در مرحله‌ی بعد فاصله‌ی تمام عناصر مجموعه داده نسبت به مرکز ثقل خوشه‌ی اول محاسبه می‌گردد. معیار فاصله‌ای که ما در الگوریتم خود لحاظ کرده‌ایم فاصله‌ی اقلیدسی است که در فرمول (۲) آمده است. در پایان این مرحله دو عنصر استخراج می‌شود. عنصر اول که کمترین فاصله با مرکز ثقل را داشته عضو جدید خوشه‌ی اول است و عنصر دوم که بیشترین فاصله را داشته اولین عنصر خوشه‌ی دوم است. پس از اضافه شدن این دو عنصر به خوشه‌های مربوطه و حذف آن‌ها از مجموعه داده‌ی اولیه مقدار مرکز ثقل این خوشه‌ها طبق فرمول (۱) به‌دست می‌آید.

$$d = \sqrt{(x_{i1}-x_{c1})^2 + (x_{i2}-x_{c2})^2 + \ldots + (x_{in}-x_{cn})^2} \qquad (۲)$$

n= تعداد صفات، i= عنصر جدید، c= مرکز ثقل خوشه

در مراحل بعدی، الگوریتم این روند را به‌طور مشابه تکرار می‌نماید، به این معنی که روند اضافه شدن دو عنصر به خوشه‌ها و حذف آن‌ها از مجموعه داده آنقدر ادامه دارد که هر یک از k خوشه دارای حداقل یک عضو باشند. با این تفاوت که فاصله‌ی نقاط تا مرکز ثقل تمامی خوشه‌های به‌دست آمده محاسبه می‌گردد. داده‌ای که بیشترین میانگین فاصله را با تمام مرکز ثقلها داشت به

[۱] Centroid

عنوان اولین عنصر خوشه‌ی جدید و داده‌ای که کمترین فاصله را با مرکز ثقل یک خوشه داشت به عنوان عضوی از آن خوشه تعیین می‌گردد. پس از اضافه شدن عناصر به خوشه‌ها به‌روز رسانی مرکز ثقل انجام می‌گیرد.

در این مرحله ما k خوشه داریم که هر کدام حداقل دارای یک عضو می‌باشند. از این پس مجموعه‌داده را جستجو کرده تا عناصر نزدیک به مرکز ثقل خوشه‌ها را بیابیم. در این جستجو در هر مرحله یک عنصر جدید یافت می‌شود. عنصر جدید از مجموعه داده حذف شده و به خوشه‌ای که کمترین فاصله را با آن دارد اضافه می‌گردد. نکته‌ی مهم این است در مرحله‌ی تعیین فاصله‌ی نقاط جدید با مرکز ثقلها فقط مرکز ثقل‌هایی شرکت دارند که خوشه‌هایشان پر نشده باشد.

تعریف (۱): خوشه‌ای پر محسوب می‌گردد که به تعداد $\lceil \frac{n}{k} \rceil$ عنصر داشته باشد که در آن n برابر تعداد عناصر مجموعه داده و k برابر تعداد خوشه‌ها می‌باشد.

به حساب نیامدن خوشه‌های پر در مرحله‌ی تعیین فاصله تضمین می‌کند که هیچ خوشه‌ای خالی نماند و تعداد اعضای نهایی خوشه‌ها حداکثر یک عنصر با هم اختلاف داشته‌باشند.

پس از آن‌که تمام عناصر مجموعه‌داده‌ی اولیه خوشه‌بندی شد، میانگین هر خوشه نیز فراهم است. از این میانگین می‌توان به عنوان مرکز ثقل ها در روش k-means سود جست. انتظار می‌رود روش مذکور برای داده‌هایی که دارای خوشه‌های دور می‌باشند بسیار خوب عمل کند.

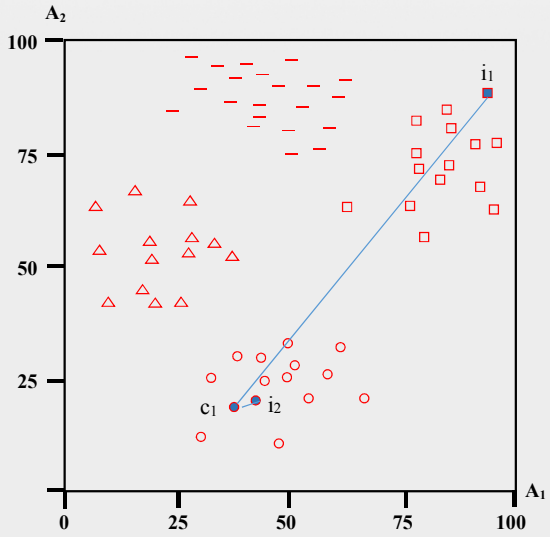
نکته‌ی جالب توجه این است که روند محاسبه‌ی مرکز ثقل‌های اولیه به دلیل ماهیت افزایشی و درنظر گرفتن کل مجموعه داده در هر مرحله، خود یک روش خوشه‌بندی محسوب می‌گردد.

این الگوریتم با فرض این‌که داده‌ها فقط شامل دو صفت نرمال شده باشند در شکل (۱) به‌تصویر کشیده شده است. شکل (۱) به‌دست آوردن تصادفی اولین مرکز خوشه c_1 را نشان می‌دهد. در روش ما فرض شده است اولین داده به عنوان اولین عضو خوشه‌ی اول بیانگر مرکز اولین خوشه است. باید توجه داشت در هر مرحله از الگوریتم مراکز خوشه جابه‌جا می‌گردند. پس از آن در دور بعدی دو داده مانند i_1 و i_2 که به‌ترتیب بیشترین و کمترین فاصله را از c_1 دارند پیدا می‌شوند. i_1 مرکز خوشه‌ی دوم و i_2 عضو بعدی خوشه‌ی اول است.

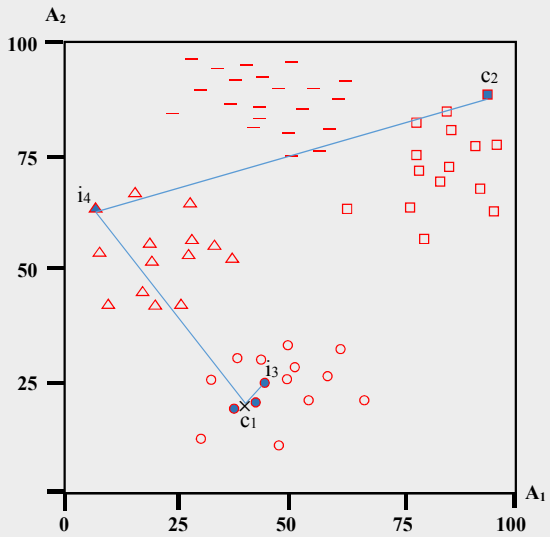
پس از اضافه شدن i_2 به خوشه‌ی اول، مقدار مرکز ثقل خوشه‌ی اول یعنی c_1 تغییر کرده و به نقطه‌ی میانگین i_1 و c_1 قبلی که با علامت × در شکل (۲) نشان داده‌ایم تغییر مکان می‌دهد. سپس تمام نقاط با تمامی مرکز ثقل‌هایی که تا کنون تعیین شده‌اند یعنی c_1 و c_2 مقایسه می‌گردند. نزدیک‌ترین نقطه به یک مرکز ثقل یعنی i_3 و نقطه‌ای که دورترین میانگین فاصله تا تمامی مرکز ثقلها را دارد یعنی i_4 یافت می‌شود. i_3 عضو جدید خوشه‌ای است که به مرکز آن نزدیک است (یعنی خوشه‌ی اول) و i_4 نیز مرکز ثقل خوشه‌ی جدید می‌باشد.

این روند تا انتخاب شدن تمامی نقاط چه به عنوان مرکز خوشه یا به عنوان عضوی از یک خوشه ادامه می‌یابد. همان‌طور که مشاهده می‌گردد از این روش می‌توان به صورت مستقل برای خوشه‌بندی سود جست.

مشکلی که در روش ما وجود دارد این است که داده‌های دور افتاده در مرحله‌ی تعیین مرکز ثقل دوم به بعد، کاندیدای خوبی به‌حساب می‌آیند. اما ماهیت این داده‌ها به‌گونه‌ای است که پس از تعیین شدن به‌عنوان مرکز ثقل، داده‌هایی نزدیک به آن یافت نمی‌شود و در واقع انحرافی برای تعیین مرکز ثقل واقعی محسوب می‌گردد. دلیل این مشکل دور بودن داده‌های دور افتاده از مرکز ثقل قبلی و تمام داده‌ها است. به واسطه‌ی دور بودن از مرکز ثقل قبلی کاندیدای بعدی برای مرکز ثقل است اما به واسطه‌ی دور بودن از بقیه‌ی داده‌ها کاندیدای خوبی برای مرکز ثقل نیست. برای رفع این مشکل می‌توان ابتدا داده‌های دور افتاده را از مجموعه داده حذف نمود.



شکل (۱): انتخاب مرکز ثقل‌های خوشه‌ی اول و دوم و عضو دوم خوشه‌ی اول (دو مرحله‌ی اول الگوریتم)



شکل (۲): انتخاب عضو سوم خوشه‌ی اول و مرکز ثقل خوشه‌ی سوم (مرحله‌ی سوم الگوریتم)

۳- معیار ارزیابی

ارزیابی روش‌های خوشه‌بندی عملی دشوار و چند معیاری است. در سال ۱۹۶۴ بونر به این نتیجه رسید که روشی جهان‌شمول برای ارزیابی یک خوشه‌بندی خوب وجود ندارد. معیارهای ارزیابی موجود به دو دسته‌ی خارجی و داخلی تقسیم می‌گردند.

معیارهای ارزیابی خارجی، خوشه‌ها را با کلاس‌های از پیش تعیین شده برای داده‌ها مقایسه می‌کنند و به دانش پیش‌زمینه‌ای نیاز دارند. روش‌های داخلی نیازی به دانش پیش‌زمینه‌ای ندارند و از دانش آماری موجود در داده‌ها و خوشه‌ها استفاده می‌کنند.

معیاری که در ارزیابی روش پیشنهادی از آن سود جسته‌ایم ضریب سیلوِات[۲] نام دارد. در این روش که در دسته روش‌های داخلی می‌گنجد ابتدا فرض می‌شود داده‌ها با روشی به k خوشه تقسیم گشته‌اند. سپس برای هر قلم داده‌ی i معیار a(i) به دست می‌آید که برابر میانگین عدم شباهت[۳] i و سایر داده‌های متعلق به خوشه‌ی دربرگیرنده‌ی i می‌باشد. در گام بعد میانگین عدم شباهت i با هر یک از خوشه‌هایی که به آن تعلق ندارد محاسبه می‌گردد و کمترین میانگین b(i) نامیده می‌شود. خوشه‌ای که کمترین شباهت را با داده‌ی i دارد خوشه‌ی مجاور i نامیده می‌شود زیرا بهترین خوشه‌ی بعدی است که i می‌تواند به آن متعلق باشد. تعریف a(i) در فرمول (۳) و b(i) در فرمول (۴) آمده است.

$$a(i)=\frac{\sum_{i'\in C_i, i\neq i'} dist(i,i')}{|C_i|-1} \qquad (۳)$$

$$b(i)=\min_{C_j:1\le j\le k, j\neq i}\left\{\frac{\sum_{i'\in C_j} dist(i,i')}{|C_j|}\right\} \qquad (۴)$$

با توجه به a(i) و b(i) می‌توان مقدار s(i) را طبق فرمول (۵) به‌دست آورد.

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \qquad (۵)$$

s(i) مقداری بین ۱ و ۱- می‌باشد. a(i) کوچک به معنی انطباق خوب داده‌ی i و خوشه‌اش می‌باشد همچنین b(i) بزرگ به معنی انطباق بد داده‌ی i با خوشه‌ی مجاورش می‌باشد. هر چه a(i) کوچکتر و b(i) بزرگ‌تر باشد s(i) به ۱ نزدیکتر بوده که نشان‌دهنده‌ی خوشه‌بندی خوب است. همچنین با منطقی مشابه s(i) نزدیک به ۱- به معنی خوشه‌بندی ضعیف می‌باشد. مقدار نهایی مورد نظر برای تعیین کیفیت یک خوشه‌بندی، میانگین کل s(i) ها می‌باشد.

٤- نتایج آزمایش

جهت نشان دادن کاربردی بودن روش پیشنهادی آن‌را پیاده‌سازی کردیم. در گام بعدی کارایی روش را برروی تعدادی مجموعه داده‌ی واقعی از جمله داده‌های گل زنبق[۴]، داده‌ی تشخیص شراب[۵] و مجموعه داده‌ی صدف[۶] آزمودیم. این سه مجموعه داده از انبار داده‌ی یادگیری ماشین UCI [۹] استخراج شده است.

مجموعه داده‌ی گل زنبق [۸] عموما به عنوان استانداردی جهت آزمون روش‌های خوشه‌بندی استفاده می‌گردد. این مجموعه داده شامل سه کلاس می‌باشد که سه گونه‌ی مختلف گل‌های زنبق از جمله (۱) زنبق ستوزا[۷]، (۲) زنبق وِرسیکالر[۸] و (۳) زنبق ویرجینیکا[۹] را نشان می‌دهد. هر یک از این کلاس‌ها دارای ۵۰ نمونه می‌باشند بنابراین مجموعه داده شامل ۱۵۰ نمونه است. هر یک از نمونه‌ها با چهار صفت مشخص شده‌اند که عبارتند از: طول کاسبرگ[۱۰]، عرض کاسبرگ، طول برگ[۱۱] و عرض برگ.

مجموعه داده‌ی شراب حاصل تحلیل شیمیایی شراب‌هایی است که در یک منطقه از ایتالیا به عمل آمده‌اند اما از سه کالتیور[۱۲] مختلف تشکیل شده‌اند. آزمایشات، مقداری از ۱۳ ماده‌ی تشکیل دهنده را در هریک این سه شراب نشان داد. این مجموعه داده شامل ۱۷۸ نمونه می‌باشد که به ترتیب ۵۹، ۷۱ و ۴۸ نمونه در کلاس‌های یک، دو و سه قرار دارند.

مجموعه داده‌ی صدف مربوط به تخمین سن صدف از طریق اندازه‌گیری‌های فیزیکی می‌باشد. تخمین سن صدف از طریق برش پوسته، رنگ آمیزی و شمارش دایره‌های آن انجام می‌گیرد. روش کارا و در دسترس دیگر برای تخمین سن صدف اندازه‌گیری معیارهایی همچون طول و وزن پوسته صدف و گوشت آن می‌باشد. این مجموعه داده شامل یک صفت کلاس به نام rings است که تعداد حلقه‌ها یا همان سن صدف را نشان می‌دهد. تعداد کلاس‌های مجموعه داده برابر ۲۸ عدد است. علاوه بر آن ۸ صفت دیگر بیان‌گر خصوصیات فیزیکی صدف می‌باشند. تعداد نمونه‌های این مجموعه داده ۴۱۷۷ عدد بوده و داده‌های غیر عددی این مجموعه داده به داده‌هایی حقیقی بین ۰ و ۱ نرمال شده است.

معیاری که برای اندازه‌گیری فاصله در نظر گرفتیم فاصله‌ی اقلیدسی است که در فرمول (۱) آمده است. تمام پیاده‌سازی‌های انجام شده به شکل ماکروهای اکسل می‌باشد. فایل اکسل ما شامل ۳ صفحه است که صفحه‌ی اول دربرگیرنده‌ی داده‌های ورودی، صفحه‌ی دوم حاوی واسط کاربر برنامه‌ی نوشته شده و صفحه‌ی سوم جهت نگهداری داده‌های موقتی است.

از طریق واسط کاربر می‌توان تعیین کرد پیش‌پردازش تصادفی بوده یا با روش پیشنهادی انجام گردد. همچنین تعداد تکرارهای k-means، مقدار k و تعداد صفات و نمونه‌هایی که برای خوشه‌بندی باید درنظر گرفته شود قابل تعیین است.

در این آزمایش تعداد خوشه‌های هر یک از مجموعه داده‌ها را برابر تعداد کلاس‌های آن مجموعه داده تعیین کرده‌ایم. برای هریک از مجموعه داده‌ها مراکز خوشه را با روش‌های تصادفی و پیشنهادی تعیین نمودیم. سپس k-means را با مراکز تعیین شده با تکرارهای مختلف انجام دادیم. معیار ارزیابی یک خوشه‌بندی خوب مقدار ضریب سیلوِات آن خوشه‌بندی است که برای هر

---

۲ Silhouette coefficient

۳ dissimilarity

۴ iris

۵ wine

۶ abalone

۷ setosa

۸ versicolor

۹ virginica

۱۰ sepal length

۱۱ petal length

۱۲ گیاه یا مجموعه‌ای از گیاهان که به جهت خصوصیات مطلوبشان انتخاب، تکثیر و نگهداری می‌شوند.

خروجی به‌دست آوردیم. نتایج آزمایشات با خوشه‌بندی اولیه تصادفی در جدول (۱) و برای خوشه‌بندی اولیه با روش پیشنهادی در جدول (۲) آمده است.

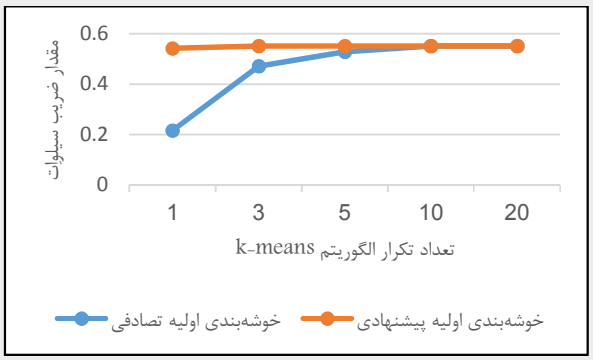
جدول ۱: مقادیر ضریب سیلوات برای خوشه‌بندی اولیه تصادفی در k-means

تکرار / داده	۱	۳	۵	۱۰	۲۰
iris	۰,۲۱۵۴۸۴۸۸۳۱۸۷۶۲۶	۰,۴۷۰۴۷۴۹۸۷۳۱۹۵۲۴	۰,۵۲۷۷۲۰۷۷۷۱۸۲۴۱۷	۰,۵۵۰۹۶۴۳۷۴۶۷۰۷۴۴	۰,۵۵۰۹۶۴۳۷۴۶۷۰۷۴۴
wine	۰,۲۱۳۵۱۸۲۱۵۴۶۵۴۵۱	۰,۴۰۵۷۴۵۸۳۷۷۲۴۵۰۵	۰,۴۱۲۵۲۴۴۳۱۸۴۶۱۰۳	۰,۴۱۶۶۴۳۶۰۳۲۶۰۲۵۲	۰,۴۱۶۶۴۳۶۰۳۲۶۰۲۵۲
abalone	۰,۲۱۴۵۵۰۲۶۸۴۷۷۶۷۹	۰,۲۶۸۱۳۵۶۶۷۲۷۵۷۷۲	۰,۲۷۴۰۴۴۰۷۴۷۲۴۹	۰,۲۸۰۱۰۶۲۶۱۲۹۵۳۸۳	۰,۲۸۴۲۵۱۹۶۷۰۵۹۳۷۸

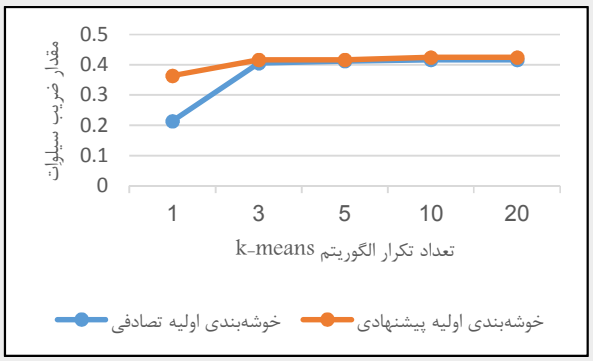
جدول ۲: مقادیر ضریب سیلوات برای خوشه‌بندی اولیه پیشنهادی در k-means

تکرار / داده	۱	۳	۵	۱۰	۲۰
iris	۰,۵۴۱۰۰۷۰۱۲۲۷۲۸۱۹	۰,۵۵۰۹۶۴۳۷۴۶۷۰۷۴۴	۰,۵۵۰۹۶۴۳۷۴۶۷۰۷۴۴	۰,۵۵۰۹۶۴۳۷۴۶۷۰۷۴۴	۰,۵۵۰۹۶۴۳۷۴۶۷۰۷۴۴
wine	۰,۳۶۴۰۱۶۶۹۹۵۸۳۲۵۳	۰,۴۱۶۶۴۳۶۰۳۲۶۰۲۵۲	۰,۴۱۶۶۴۳۶۰۳۲۶۰۲۵۲	۰,۴۲۴۶۹۷۰۲۳۵۸۷۳۷۱	۰,۴۲۴۶۹۷۰۲۳۵۸۷۳۷۱
Abalone	۰,۲۴۴۲۸۴۲۴۹۸۰۴۹۰۴	۰,۳۳۴۲۳۶۹۵۷۴۷۰۶۹۵	۰,۳۴۹۳۸۹۵۹۰۷۷۶۸۵۵	۰,۳۶۱۱۲۵۱۷۱۲۵۵۷۶۵	۰,۳۶۱۱۲۵۱۷۱۲۵۵۷۶۵

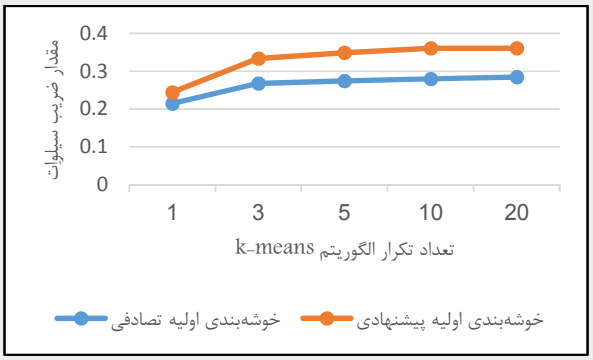
در شکل (۳) سرعت هم‌گرایی الگوریتم k-means با خوشه‌بندی اولیه پیشنهادی و تصادفی برای مجموعه داده‌ی گل زنبق مقایسه گشته است. همچنین در شکل‌های (۴) و (۵) این نمودار برای مجموعه داده‌ی شراب و صدف نشان داده شده است.



شکل ۳: مقایسه‌ی سرعت همگرایی الگوریتم k-means با خوشه‌بندی‌های اولیه‌ی تصادفی و پیشنهادی برای داده‌ی زنبق



شکل ٤: مقایسه‌ی سرعت همگرایی الگوریتم k-means با خوشه‌بندی‌های اولیه‌ی تصادفی و پیشنهادی برای داده‌ی شراب



شکل ۵: مقایسه‌ی سرعت همگرایی الگوریتم k-means با خوشه‌بندی‌های اولیه‌ی تصادفی و پیشنهادی برای داده‌ی صدف

۵- نتیجه‌گیری

با نگاهی به جدول‌های (۱) و (۲) مشاهده می‌گردد سرعت همگرایی k-means با خوشه‌بندی اولیه تصادفی از روش پیشنهادی به مراتب کمتر می‌باشد. نکته مهم دیگر این است که خوشه‌بندی اولیه ارائه شده همیشه یک نتیجه دربر خواهد داشت در حالیکه خوشه‌بندی اولیه تصادفی در هر اجرا نتیجه‌ی متفاوتی دربر دارد.

به نظر می‌رسد روش پیشنهادی برای خوشه‌بندی اولیه انقدر خوب باشد که بتوان بدون نیاز به k-means از آن به‌عنوان خوشه‌بندی نهایی سود جست. به این منظور می‌توان صفات دخیل در تعیین خوشه‌ها را با دقت بیشتر تعیین نمود. همچنین می‌توان نقطه‌ی اولیه را با روشی دیگر همچون میانگین‌گیری از نقاط تعیین کرد.

روش پیشنهادی به داده‌های دور افتاده حساس می‌باشد. انتظار می‌رود با حذف داده‌های دور افتاده کیفیت خوشه‌ها بهبود یابد.

٦- منابع

[1] Mac Queen, J., 1967. Some methods for classification and analysis of multivariate observations (pp. 281297). In: Le Cam, L.M., Neyman, J. (Eds.), Proc. 5th Berkley Symp. on Mathematical Statistics and Probability, vol. 1. University of California Press, p. 666, xvii.

[2] Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis. John Wiley and Sons, NY.
[3] Jain, A.K., Dubes, R.C., 1988. Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ.
[4] Bradley, P.S., Fayyad, U.M., 1998. Refinning initial points fork-means algorithm. In: Proceeding of the 15th Internat. Conf. on Machine Learning (ICML'98).
[5] Likas, A., Vlassis, N., Jakob, J.V., 2003. The globalk-means algorithm algorithm. Pattern Recognition 36, 451–461.
[6] Khan, S.S., Ahmad, A., 2004. Cluster center initialization algorithm fork-means algorithm. Pattern Recognition Lett. 25, 1293–1302.
[7] Deelers, S., Auwatanamongkol, S., 2007. Enhancingk-means algorithm with initial cluster centers derived from data partitioning along the data axis with the highest variance. Internat. J. Comput. Sci. 2, 247–252.
[8] Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugenic. 7 (part 2), 179–188.
[9] http://archive.ics.uci.edu/ml/datasets.html