dataacademy.ir

خوشه بندی داده‌ها به روش الگوریتم فاخته

مهری ملالو[5] و محمد ابراهیم شیری[1]
[5] دانشجوی کارشناسی ارشد علوم کامپیوتر، دانشگاه صنعتی امیرکبیر، m.mollalo@aut.ac.ir
[1] عضو هیئت علمی گروه علوم کامپیوتر، دانشگاه صنعتی امیرکبیر ، shiri@aut.ac.ir

چکیده - خوشه‌بندی از جمله روش‌های پرکاربرد در تجزیه و تحلیل داده‌ها است که بدون هیچ دانش قبلی، داده‌ها را به گروه‌های معنی‌داری تقسیم می‌کند که یکی از موضوعات اساسی در داده کاوی است. در این مقاله یک الگوریتم خوشه‌بندی جدید بر اساس سبک خاص زندگی جمعی یک دسته پرنده بنام فاخته ارائه شده است. الگوریتم فاخته که یکی از روش‌های جدید بهینه‌سازی تکاملی می‌باشد که با الهام گرفتن از روش تخم‌گذاری فاخته‌ها، پرورش تخم‌ها، تلاش برای زنده ماندن در میان دیگر فاخته‌ها و مهاجرت به سمت محیط بهتر برای مساله خوشه‌بندی داده‌ها استفاده شده است. الگوریتم معروف خوشه‌بندی *k-means* به مقداردهی اولیه بسیار حساس است و به راحتی در بهینه‌های محلی گیر می‌افتد در صورتی که الگوریتم پیشنهاد شده به خوشه‌های با اندازه و ابعاد متفاوت حساس نیست، برای مجموعه داده‌های چند بعدی مناسب است و همچنین سرعت بسیار خوبی در همگرایی به نقطه بهینه سراسری و دقت بالا در حل مسائل از خود نشان داده است. کارایی الگوریتم پیشنهادی بر روی مجموعه‌های داده‌ای پایگاه UCI مورد ارزیابی قرار گرفته شده است. نتایج بدست آمده از آزمایشات نشان دهنده بهبود عملکرد این الگوریتم در مقایسه با الگوریتم‌های معروف خوشه‌بندی همچون k-means و PSO است.

کلید واژه- الگوریتم بهینه‌سازی فاخته، الگوریتم تغییریافته فاخته، خوشه بندی، داده کاوی

1- مقدمه

خوشه بندی، الگوها و نمونه‌ها را بدون دخالت هیچ ناظری، در گروه‌های مختلفی دسته‌بندی می‌کند[5]. در پنجاه سال گذشته از نقطه نظرهای زیادی به مسئله خوشه بندی توجه شده است. این مسئله در زمینه‌های مختلف از قبیل تشخیص الگو، تجزیه و تحلیل داده‌ها، پردازش تصویر، علوم اقتصادی، زیست-شناسی، بیوانفورماتیک و فشرده‌سازی داده‌ها مورد استفاده قرار گرفته شده است. بنابراین مطالعه در مورد الگوریتم‌های جدید خوشه‌بندی در زمینه‌های تحقیقاتی از جمله داده‌کاوی، یادگیری‌ماشین، آمار، و زیست‌شناسی موضوع مهمی است.

هدف اصلی روش‌های ارائه شده برای خوشه‌بندی، پیدا کردن گروه‌هایی از داده‌ها است که دارای بیشترین شباهت بین داده‌های درون هر خوشه و کمترین شباهت بین داده‌های بین هر دو خوشه است. در سال‌های اخیر الگوریتم‌های خوشه‌بندی زیادی از جمله، سلسه مراتبی[1] ، الگوریتم‌های بخش‌بندی[9-4]، الگوریتم‌های مبتنی بر چگالی[1] ، الگوریتم‌های مبتنی بر مدل [6] و الگوریتم‌های مبتنی بر مش[7] ارائه شده است.

دربسیاری از روش‌های خوشه‌بندی، فاصله (میزان شباهت) داده‌های متعلق به یک خوشه نسبت به فاصله آنها با داده‌های دیگر خوشه‌ها، معیار تعیین خوشه در نظر گرفته شده است. در ساده‌ترین روش، هر داده می‌تواند با در نظر گرفتن میزان شباهت یا کمترین فاصله‌اش تا خوشه‌ها، دقیقا در یکی از خوشه‌ها عضویت داشته باشد. [8-9].

به غیر از روش‌های فوق، الگوریتم خوشه‌بندی k-means برای شناسایی خوشه‌هایی که در فضای ورودیِ غیر خطیِ از هم جدا هستند، استفاده می‌شود[51-51].

الگوریتم k-means [9] به دلیل سادگی پیاده سازی و هزینه کم محاسباتی و کارایی بالا، محبوب‌ترین روش خوشه‌بندی داده-ها است. با این حال یکی از مشکلات این الگوریتم این است که به مقداردهی اولیه بسیار حساس است و به راحتی در بهینه‌های محلی گیر می‌افتد. بطوریکه با افزایش ابعاد مجموعه داده‌ها، پیدا کردن جواب بهینه به یک مساله NP-hard تبدیل می‌شود[59]. لذا به دلیل اهمیت استراتژی خوشه‌بندی، جدیداً از الگوریتم‌های بهینه‌سازی سراسری تصادفی همچون الگوریتم ژنتیک(GA)،

dataacademy.ir

الگوریتم کلونی مورچه‌ها(ACO) و الگوریتم اجتماع ذرات(PSO) برای خوشه‌بندی داده‌ها استفاده شده است[54-57] در همه این روش‌ها، الگوریتم با یک جمعیت اولیه شروع بکار کرده و سپس با تعدادی تکرار، فضای راه حل را برای رسیدن به جواب تقریباً بهینه جستجو می‌کند.

بیشتر الگوریتم‌های ارائه شده برای حل مسائل بهینه‌سازی، از رفتار حشرات و پرندگان در طبیعت الهام گرفته شده‌اند. الگوریتم بهینه‌سازی فاخته (COA[1]) یک الگوریتم بهینه‌سازی تکاملی می‌باشد که از سبک خاص زندگی یک نوع پرنده بنام فاخته[2] الهام گرفته شده است. تخم‌گذاری و تولیدمثل خاص فاخته‌ها اساس این الگوریتم بهینه‌سازی می‌باشد. در این الگوریتم فاخته‌ها در دو مدل استفاده شده‌اند: فاخته‌های بالغ و فاخته‌هایی که در تخم هستند. فاخته‌های بالغ در لانه‌ی بعضی پرندگان دیگر تخم‌گذاری می‌کنند و اگر این تخم‌ها توسط پرندگان میزبان شناسایی نشوند و از بین نروند، رشد کرده و به فاخته‌ی بالغ تبدیل می‌شوند. ویژگی‌های محیطی و مهاجرت گروه‌های فاخته‌ها، باعث همگرا شدن الگوریتم و پیدا کردن جواب بهینه سراسری می‌شود. در این مقاله الگوریتم بهینه‌سازی فاخته برای خوشه‌بندی داده‌ها گسترش داده شده است و کارایی آن با الگوریتم‌هایی همچون k-means و PSO بر روی مجموعه‌های داده‌ای مختلفی از پایگاه داده UCI مورد مقایسه قرار گرفته شده است. نتایج بدست آمده، نشان دهنده کارایی عملکرد این روش نسبت به روش‌های دیگر خوشه‌بندی می‌باشد.

در ادامه در بخش 1 مساله خوشه‌بندی داده‌ها را مطرح می‌شود. سپس در بخش 9 ویژگی زندگی پرنده فاخته و جزئیات مربوط به تخم‌گذاری و مهاجرت آن و همچنین الگوریتم پیشنهاد شده با جزییات بیشتری مورد بررسی قرار خواهد گرفت. در ادامه در بخش 4 نتایج بدست آمده از شبیه‌سازی الگوریتم پیشنهاد شده بر روی مجموعه‌های داده‌ای نشان داده شده است. در نهایت در بخش 1 نتیجه‌گیری کارها ارائه شده است.

2- خوشه‌بندی

هدف از انجام خوشه‌بندی، طبقه‌بندی اشیا مطابق با میزان تشابه بین آنها و سازماندهی داده‌ها در دو یا چند گروه است. تکنیک‌های خوشه‌بندی از جمله روش‌های بدون نظارت است و هیچ شناخت اولیه‌ای از کلاس‌ها و یا در واقع خوشه‌ها در آن وجود ندارد.

ایده‌ی خوشه‌بندی برای اولین بار در دهه‌ی 5991 ارائه شد و امروزه مورد توجه بسیاری از محققان قرار گرفته است. خوشه به یک گروه از اشیا اطلاق می‌شود که شباهت بیشتری با هم نسبت به اشیا در خوشه‌های دیگر دارند. عبارت " تشابه" باید به معنای تشابه ریاضی و قابل محاسبه تفسیر گردد.

در فضای داده‌ای، تشابه دو شیء معمولاً به معنای فاصله آن دو شیء در فضای برداری بیان می‌شود. فاصله می‌تواند در بین بردار داده‌ها با هم و یا فاصله‌ی یک داده تا یک نمونه‌ی معرف محاسبه شود. نمونه‌ی معرف معمولاً از پیش شناخته شده نیست و در طی انجام خوشه‌بندی و جداسازی داده‌ها تعیین می‌شود.

مسئله‌ی خوشه‌بندی را بصورت کلی می‌توان به شرح زیر بیان کرد: در مجموعه داده‌ی $X=\{x_1,x_2,\ldots,x_N\}$ ، N شی داده شده به K خوشه $(C_1,C_2,\ldots,C_K)$ تقسیم می‌شود. بطوریکه، شباهت بین داده‌های درون هر خوشه حداکثر و شباهت بین داده‌های بین خوشه‌های متفاوت حداقل شود. این مساله در معادله (5) نمایش داده شده است.

$$\begin{cases} \bigcup_{i=1}^{K} C_i = X \\ C_i \cap C_j = \phi \qquad i,j=1,2,\ldots,K \qquad i \neq j \\ C_i \neq \phi i=1,2,\ldots,K \qquad i=1,2,\ldots,K \end{cases} \tag{5}$$

که K تعداد خوشه‌ها و C_i خوشه i ام می‌باشد.

از نقطه نظر ریاضی خوشه ی C_i می‌تواند به صورت معادله(1) مشخص شود:

$$\begin{cases} C_i = \left\{ x_j \,\middle|\, \|x_j - z_i\| \le \|x_j - z_p\|, x_j \in X \right\}, p \neq i, p=1,\ldots,K \\ z_i = \dfrac{1}{|C_i|} \sum_{x_j \in C_i} x_j \qquad i=1,2,\ldots,K \end{cases} \tag{1}$$

$\| \ \|$ بیانگر فاصله اقلیدسی یا فاصله کسینوسی بین هر دونقطه در مجموعه نمونه است. z_i مرکز خوشه C_i است که در واقع میانگین تمامی نقطه‌ها در خوشه را نشان می‌دهد.

[1] Cuckoo Optimization Algorithm
[2] Cuckoo

dataacademy.ir

از معادله (1) در می‌یابیم که C_i توسط داده‌هایی که نزدیکترین عناصر به z_i هستند ساخته می‌شود. بنابراین وظیفه خوشه‌بندی انجام فرایندی است که K مرکز خوشه $(C_1, C_2, \ldots, C_K)$ را تعیین کند. به طوریکه N شی داده شده به K خوشه تقسیم‌بندی می‌شد. ابتدا هر یک از اشیا به یک خوشه اختصاص داده می‌شود به طوریکه مجموع مجذور فاصله اقلیدسی بین هر شی و مرکز خوشه‌ای که به آن شی اختصاص داده شده است حداقل باشد. این مساله در معادله (9) نمایش داده شده است که جمع مربع خطاها (SSE[3]) می‌باشد[58].

$$SSE = \sum_{i=1}^{K} \sum_{x_j \in c_i} \left\| x_j - z_i \right\|^2 \qquad (9)$$

که $z_i = \{z_1, z_2, \ldots, z_K\}$ نشان دهنده K مرکز خوشه است و $x_j \in \{x_1, x_2, \ldots, x_N\}$ و N تعداد داده‌های مجموعه داده‌ای می‌باشد.

برای هر یک از داده ها در مجموعه ارائه شده، میزان فاصله تا نزدیکترین خوشه، خطا می باشد.

3- روش پیشنهاد شده: خوشه‌بندی داده‌ها به روش الگوریتم فاخته

3-1- شیوه‌ی خاص زندگی و تخمگذاری فاخته‌ها

فاخته‌ها پرندگانی هستند که خود را از دردسر هرگونه لانه‌سازی و وظایف والدین رهانیده‌اند و به نوعی زیرکی جهت پرورش جوجه‌های خود متوسل شده‌اند. این پرندگان تخم‌های خود را در لانه سایر انواع پرندگان قرار می‌دهند و صبر می‌کنند تا آنها در کنار تخم‌های خود به تخم‌های این پرندگان نیز رسیدگی کنند. در این بین هستند پرندگانی که تخم‌های فاخته‌ها را در لانه‌های خود تشخیص می‌دهند و حتی بعضا تخم‌های فاخته را از لانه بیرون پرت می‌کنند[59].

جوجه‌های فاخته زودتر از تخم‌های پرنده میزبان از تخم بیرون می‌آیند و زودتر هم رشد می‌کنند. در اکثر موارد جوجه فاخته تخم‌ها و یا جوجه‌های پرنده میزبان را از لانه بیرون می‌اندازند و در نهایت در هر لانه یک تخم شانس رشد را پیدا می‌کند.

3-2- الگوریتم پیشنهاد شده جهت خوشه بندی داده‌ها

روش پیشنهاد شده جهت خوشه‌بندی داده‌ها به روش الگوریتم فاخته (CCA[4]) برای مسال با ابعاد زیاد کارایی خوبی از خود نشان داده است. همانند سایر الگوریتم‌های تکاملی،CCA هم با یک جمعیت اولیه کار خود را شروع می‌کند.

این جمعیت از فاخته‌ها هر کدام بدنبال پیدا کردن K مرکز خوشه بهینه می‌باشند که برای رسیدن به چنین هدفی باید تولید مثل و تخم‌گذاری کنند تا محیط توسط فاخته‌های بیشتری جستجو شود. فاخته‌ها تخم‌های خود را در لانه تعدادی پرنده‌ی میزبان قرار می‌دهند. تعدادی از تخم‌ها توسط پرنده میزبان شناسایی شده و از بین می‌روند و برخی دیگر شانس زنده ماندن خواهند داشت. موقعیتی که در آن بیشترین تعداد تخم‌ها نجات یابند پارامتری خواهد بود که COA قصد بهینه‌سازی آن را دارد.

مقدار تابع برازش برای مسائل خوشه‌بندی داده‌ها برابر میانگین فاصله داده‌ها از مراکز خوشه‌ها می‌باشد و به صورت معادله (4) تعریف می‌شود:

$$fitness(z_i) = \left(\frac{1}{n} \sum_{m=1}^{K} \sum_{x_j \in c_m} \left\| x_j - z_{im} \right\| \right)^{-1} \qquad (4)$$

که $z_i = \{z_{i1}, z_{i2}, \ldots, z_{iK}\}$ نشان دهنده فاخته i ام است که شامل K مرکز خوشه است و z_{im} نمایانگر مرکز خوشه m ام از فاخته i ام است. همچنین $x_j \in \{x_1, x_2, \ldots, x_N\}$ و N تعداد داده‌های مجموعه داده‌ای می‌باشد. مقدار تابع برازش فوق نشان می‌دهد که مراکز خوشه با میانگین کمترین فاصله از مقدار برازش بیشتری برخوردارند

در طبیعت هر فاخته بین 1 تا 11 تخم می‌گذارد. عادت دیگر هر فاخته این است که در یک دامنه مشخص تخم‌های خود را می‌گذارد که حداکثر دامنه تخمگذاری (ELR[5]) می‌باشد. در

[3] Sum of squared error

[4] Cuckoo Clustering Algorithm

[5] Egg Laying Radius

dataacademy.ir

یک مساله بهینه‌سازی به حد بالای متغیرها var_{hi} و به حد پایین متغییرها var_{low} می‌گویند. هر فاخته دارای ELR مخصوص به خود خواهد بود که با تعداد کل تخم‌های موجود در محیط، تعداد تخم‌های فعلی فاخته و همچنین حد بالا و پایین متغیرهای مساله متناسب است.

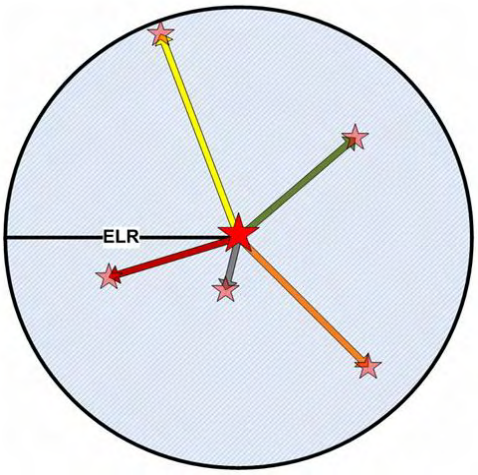
ELR به صورت معادله(1) محاسبه می‌شود:

$$ELR = \alpha \times \frac{\text{Number of current eggs}}{\text{Total number of eggs}} \times (var_{hi} - var_{low}) \quad (1)$$

α متغیری است که حداکثر شعاع تخم‌گذاری با آن تنظیم می‌شود.

هر فاخته بصورت تصادفی تخم‌هایی را در لانه پرندگان میزبان که در ELR خود قرار دارد، می‌گذارد (شکل 1).

نکته جالب در مورد جوجه فاخته‌ها این است که فقط یک تخم در هر لانه امکان رشد دارد. چرا که وقتی جوجه‌های فاخته از تخم بیرون می‌آیند تخم‌های پرنده میزبان را از لانه بیرون می‌کنند و اگر جوجه‌های پرنده میزبان زودتر از تخم خارج شده باشند جوجه فاخته بیشترین مقدار غذا را که پرنده میزبان می‌آورد، خواهد خورد و پس از چند روز جوجه‌های خود پرنده میزبان از گرسنگی می‌میرند و فقط جوجه فاخته زنده می‌ماند.



شکل 1. تخمگذاری تصادفی در محدوده‌ی ELR، ستاره قرمز رنگی که در وسط قرار دارد محل سکونت اولیه‌ی فاخته‌ای با 5 تخم‌مرغ می‌باشد؛ ستاره‌های صورتی لانه‌ی جدید تخم‌مرغ‌ها هستند.

وقتی جوجه فاخته‌ها رشد کردند و تبدیل به فاخته بالغ شدند، پرنده‌های فاخته دیگر با میزانی فاصله و انحراف به سمت بهترین منطقه فعلی در بین تمام فاخته‌ها که در آنجا شانس زنده ماندن تخم‌ها بیشتر است مهاجرت می‌کنند.

هنگام مهاجرت به سمت نقطه هدف، فاخته‌ها تمام مسیر را به سمت محل هدف طی نمی‌کنند. آنها فقط قسمتی از مسیر را طی کرده و در آن مسیر نیز مقداری انحراف دارند. این نحوه حرکت در شکل 1 بوضوح قابل مشاهده است.

همانطور که در شکل 2 دیده می‌شود هر فاخته فقط درصدی از کل مسیر را به سمت هدف ایده‌آل فعلی طی می‌کند($\lambda\%$) و یک انحراف φ نیز دارد. این دو پارامتر به فاخته‌ها کمک می‌کنند تا محیط بیشتری را جستجو کنند و از گیر افتادن در بهینه‌های محلی اجتناب کنند. λ عددی تصادفی بین 0 و 1 است و φ عددی بین $\pi/6$ و $-\pi/6$ می‌باشد.

وقتی تمام فاخته‌ها به سمت نقطه هدف مهاجرت کردند و نقاط سکونت جدید هرکدام مشخص شد، دوباره مراحل تخمگذاری و محاسبه ELR برای هر یک از فاخته‌ها انجام می‌شود.

با توجه به این واقعیت که همیشه تعادلی بین جمعیت پرندگان در طبیعت وجود دارد عددی مثل Nmax حداکثر تعداد فاخته هایی که می توانند در یک محیط زندگی کنند را کنترل می‌کند. این تعادل بدلیل محدودیت‌های غذایی، شکار شدن توسط شکارچیان و نیز عدم امکان پیدا کردن لانه های مناسب برای تخم‌ها وجود دارد.

شکل2. مهاجرت یک نمونه فاخته به سمت محل اقامت هدف

شبه کد الگوریتم خوشه‌بندی داده‌ها به روش زندگی فاخته (CCA) به صورت زیر می‌باشد (الگوریتم1):

الگوریتم 1: شبه کد الگوریتم خوشه‌بندی داده‌ها به روش الگوریتم (CCA)

Require: Data set, $X = \{x_1, x_2, \ldots, x_N\}$;
Cluster number, K;
Ensure: Clusters: $\{C_1, C_2, \ldots, C_K\}$;

dataacademy.ir

Initialize:
numCuckooS: number of initial population;
minNumberOfEggs: minimum number of eggs for each cuckoo;
maxNumberOfEggs: maximum number of eggs for each cuckoo;
maxIter: maximum iteration of the Cuckoo Clustering Algorithm;
maxNumOfCuckoos: maximum number of cuckoos that can live at the same time;
varlow;
varhi;
Generate the Position of Initialize Population randomly which each single $z = \{z_1, z_2, \ldots, z_{numCuckooS}\}$; cuckoo
contains **K** randomly generated centroid vectors z_i
$z_i = \{z_{i1}, z_{i2}, \ldots, z_{iK}\}$
for t=1:maxIter **do**
  for i=1:numCuckooS **do**
    Initialize number of eggs for each z_i
  end for
  for i=1:numCuckooS **do**
    Calculate ELR(egg laying radius) according to Eq.(5) for current z_i
  end for
  Lay eggs in different nests
  Remove the eggs on the same positions for all Cuckoos
  for i=1:numCuckooS **do**
    Evaluate fitness function for each z_i **and** his eggs and put them under each other
  end for
  if Population > maxNumOfCuckooS **then**
    Kill cuckoos in worst position
  end if
  Update the global best position $zbest_i$: Select the best $zbest_i$ from $\{z_1, z_2, \ldots, z_{numCuckooS}\}$
  Move all cuckoos toward $zbest_i$ (Immigration of cuckoo)
end for

4- نتایج آزمایشات

4-1- مجموعه داده

در این بخش نتایج شبیه‌سازی روش پیشنهادی خوشه‌بندی به روش الگوریتم فاخته (CCA) نسبت به روش‌های k-means و PSO ارائه داده می‌شود.

در این مقاله، برای بررسی و ارزیابی صحت عملکرد الگوریتم پیشنهادی از سه مجموعه Iris، Wine و Glass از مجموعه داده‌های مربوط به پایگاه داده UCI استفاده شده است. جزئیات مربوط به این مجموعه داده‌ها در جدول 5 ارائه شده است.

الگوریتم پیشنهادی بر روی سه مجموعه داده‌ای ارزیابی شده است. مجموعه داده‌ای Iris شامل داده‌هایی در مورد سه نوع گل زنبق است که دارای 511 نمونه داده است که به سه کلاس تقسیم شده‌اند که در هر کلاس 11 نمونه قرار دارد و هر نمونه چهار متغیر ورودی دارد. مجموعه داده‌ای Wine دارای 578 نمونه داده است که به سه دسته یا کلاس با 59 متغیر و بعد تقسیم شده است. مجموعه داده‌ای Glass یکی دیگر از مجموعه‌هاست که در رابطه با اجزاء شیمیایی تشکیا دهنده 6 نوع شیشه است و دارای 6 کلاس است و از 154 نمونه تشکیل شده است و همچنین دارای 9 فیلد و ورودی است.

جدول 5: مشخصات مجموعه‌های داده‌ای

مجموعه داده	نمونه‌ها	ویژگی/ بعد داده‌ها	خوشه
Iris	511	4	9
Wine	578	59	9
Glass	154	9	6

4-2- تنظیم پارامترها

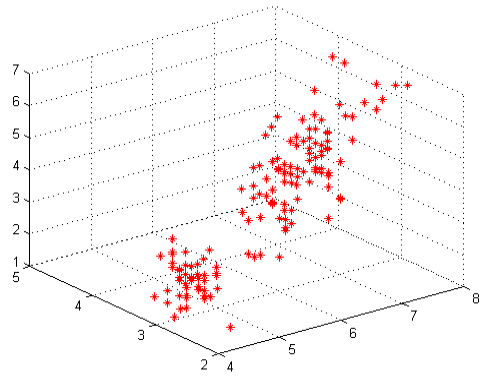
پارامترهای الگوریتم‌های خوشه بندی مختلف به این صورت تنظیم شده‌اند که برای PSO، 11 ذره اولیه و $w = 0.72$ و $C_1 = C_2 = 1.49$ را تنظیم شده است.

برای CCA جمعیت اولیه 11 فاخته و حداکثر فاخته‌هایی که می‌توانند در محیط زنده باشند نیز 11 در نظر گرفته شده است. حداقل و حداکثر مقادیر تخمی که هر فاخته می تواند 9 و 8 است.

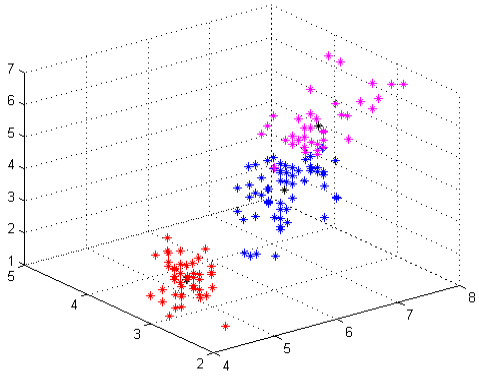
4-3- نتایج شبیه‌سازی

در این آزمایشات، تعداد تکرارهای الگوریتم برابر 91 تکرار در نظر گرفته شده است و از معیار فاصله اقلیدسی برای به دست آوردن فاصله بین نمونه‌ها نسبت به مراکز خوشه‌ها استفاده شده است. شکل 9 مجموعه‌داده‌ای Iris را قبل از خوشه‌بندی نشان می‌دهد و در شکل 4 مجموعه داده‌ای Iris بعد از خوشه‌بندی نمایش داده شده است. این مجموعه داده‌ای دارای چهار بعد است. همانطور که به روشنی در تصویر قابل مشاهده است، این الگوریتم به درستی، مجموعه نمونه‌های ورودی را به سه خوشه و دسته تقسیم‌بندی کرده است.

dataacademy.ir



شکل9: مجموعه داده‌ای Iris قبل از خوشه‌بندی



شکل 4: مجموعه داده‌ای Iris بعد از خوشه‌بندی که 9 خوشه موجود با رنگ‌های متفاوت نشان داده شده‌اند.

نتایج شبیه سازی‌های انجام شده بر روی مجموعه‌های داده‌ای Iris، Wine و Glass به صورت میانگین بعد از 91 مرتبه تکرار الگوریتم، در جدول (1) ارائه شده است. یکی از ویژگی‌های الگوریتم پیشنهادی، سرعت قابل توجه آن در همگرا شدن به جواب نزدیک به جواب بهینه است. این الگوریتم در تکرارهای[6] پایین (تا تکرار 51) به جواب نزدیک به جواب بهینه همگرا می‌شود. همانطور که در تصویر 1 قابل مشاهده است این الگوریتم در هر 91 تکراری که انجام شده است به صورت میانگین تا تکرار 51 به جواب تقریباً بهینه همگرا شده است که این نشان از سرعت بالای الگوریتم در رسیدن به جواب در تکرارهای پایین می‌باشد. الگوریتم پیشنهاد شده (CCA) با الگوریتم‌های خوشه‌بندی k-means و PSO بر اساس دو معیار دقت[7] و میانگین فاصله نمونه‌ها از مراکز، با هم مقایسه شده‌اند.

معیار دقت با استفاده از برچسب کلاس‌ها، میزان دقت روش خوشه‌بندی در انتساب داده‌ها به خوشه‌ها را نشان می‌دهد. لازم به ذکر است که برچسب کلاس‌ها تنها برای ارزیابی نتایج استفاده می‌شوند و در فرآیند خوشه‌بندی دخالتی ندارند. رابطه (6) این معیار را نشان می‌دهد.

$$\text{Accuracy}(\pi) = \frac{\sum_{m=1}^{p} majority(C_m | L_m)}{N} \qquad (6)$$

$majority(C_m | L_m)$ در رابطه (6) تعداد داده‌ای هایی از یک کلاس است که دارای اکثریت در خوشه C_m می‌باشند.

نتایج حاصل از شبیه‌سازی الگوریتم‌ها بر روی مجموعه‌های داده‌ای Iris، Wine و Glass در جدول 1 مشاهده می‌شود.

با مقایسه نتایج گزارش شده در جدول 1 مشاهده می‌شود که الگوریتم پیشنهادی، نسبت به الگوریتم‌های خوشه‌بندی k-means و PSO از دقت و کارایی بالاتری برخوردار است.

جدول 1: مقایسه کارایی الگوریتم‌های k-means، PSO و CCA بر روی مجموعه‌های داده‌ای موجود در جدول5

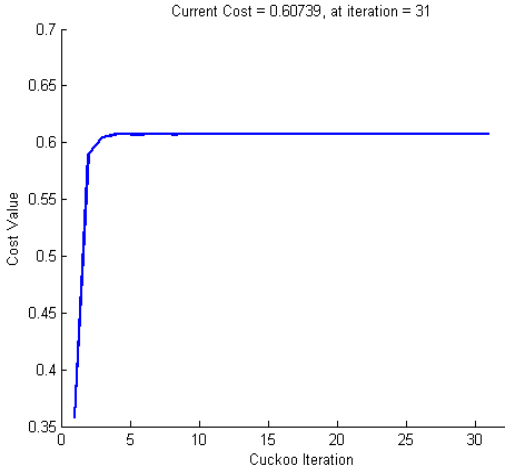
مجموعه داده	روش	دقت	میانگین فاصله تا مرکز
Iris	k-means	1/881	97/9461
	PSO	1/999	97/9119
	CCA	**0/998**	**97/0510**
Wine	k-means	1/711	5661/6
	PSO	1/699	5614/8
	CCA	**0/776**	**1627/4**
Glass	k-means	1/477	141/1749
	PSO	1/481	198/9111
	CCA	**0/508**	**213/5036**

همانطور که در مجموعه داده‌ای Iris مشاهده می‌شود، الگوریتم CCA میانگین فاصله تا مرکز را 97/1151 تشخیص داده است و این در صورتی است که الگوریتم‌های k-means و PSO به ترتیب مقادیر 97/9461 و 97/9119 را برای این معیار تشخیص داده‌اند که با مقایسه این مقادیر، برتری روش CCA نسبت به الگوریتم‌های k-means و PSO قابل مشاهده است. زیرا هرچه مقدار بدست آمده از معیار میانگین فاصله تا مرکز کمتر و مقدار معیار دقت بزرگتر باشد نشان دهنده عملکرد بهتر الگوریتم

[6] Iterations
[7] Accuracy

dataacademy.ir

در خوشه‌بندی داده‌ها می‌باشد.



تصویر5. همگرا شدن الگوریتم به جواب تقریباً بهینه در تکرارهای پایین

5- نتیجه‌گیری

در این مقاله یک الگوریتم جدید، مبتنی بر سبک خاص زندگی پرندگان فاخته جهت خوشه‌بندی داده‌ها ارائه شد. در این الگوریتم از روش تخم‌گذاری و مهاجرت فاخته‌ها برای خوشه-بندی داده‌ها الهام گرفته شد. خوشه‌بندی داده‌ها در واقع پیدا کردن بهینه‌ترین مرکز خوشه‌ها می‌باشد که این کار بوسیله‌ی بهینه کردن تابع برازش انجام می‌شود. یکی از مزایای این الگوریتم این است که به راحتی از گیرافتان در بهینه‌های محلی اجتناب می‌کند و همچنین سرعت اجرای الگوریتم نسبت به الگوریتم‌های خوشه‌بندی PSO و k-means مقایسه شد. نتایج شبیه‌سازی‌های انجام شده بر روی مجموعه‌های داده‌ای مختلف نشان از بهبودکارایی و عملکرد CCA در مقایسه با الگوریتم‌های خوشه‌بندی k-means و PSO است.

سپاسگزاری

در این قسمت جا دارد که از استاد فرزانه و دانشمند، **جناب آقای دکتر پرهام مرادی دولت آبادی**، به پاس حمایت‌ها و راهنمایی‌های ارزنده‌شان تشکر و قدردانی کنم و برای ایشان از خداوند منان عزت روز افزون را خواستارم.

مراجع

[1] A.K. Jain, M.N. Mutry, P.J. Flyn, Data clustering: a review, ACM Computing Surveys
[2] S. Guha, et al., “CURE: an efficient clustering algorithm for large databases,” SIGMOD Rec., vol. 27, pp.73_84, 1998.
[3] J.MacQueen, “Some Methods for Classificarion and Analysis of Multivariate Observations,” presented at the Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability- Vol. 1, 1967.
[4] R. T. Ng and J. Han, “Efficient and Effective Clustering Methods for Spatial Data Mining,” presented at the Proceedings of the 20th International Conference on Very Larg Data Bases, 1994.
[5] A. Hinneburg, et al., “An Efficient Approach to Clustering in Large Multimedia Databases with Noise,” 1998, pp.58_65.
[6] A. P. Dempster, et al., “Maximum Likelihood from Incomplete Data via the EM Algorithm,” Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, pp. 1_38, 1977.
[7] G. Sheikholeslami, et al., “WaveCluster: a wavelet-based clustering approach for spatial data in very large databases,” The VLDM Journal, vol. 8, pp. 289_304, 2000.
[8] Duda, R., Hart, P., (1973), Pattern Classification and Scene Analysis. J. Wiley & Sons, Inc. New York, NY, USA.
[9] Krishnapuram, R. and Keller, J., (1996), The possibilistic c-means algorithm: insightes and recommendations. IEEE Trans. Fuzzy Systems 4,385-393.
[10] I.S. Dhillon, Y. Guan, B. Kulis, Weighted graph cuts without eigen vectors: a multilevel approach, IEEE Transactions on Pattern Analysis Machin Intelligence 29 (11) (2007) 1944_1957.
[11] I.S. Dhillon, Y. Guan, B. Kulis, A unified view of kernel k-means, spectral clustering and graph partitioning, Technival Report TR_04_25, UTCS, 2005.
[12] I.S. Dhillon, Y. Guan, B. Kulis, Kernel k-means: spectral clustering and normalized cuts, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, NY, USA, 2004, pp.551_556.Sdfsfsf
[13] X.Rui and D. Wunsch, II, “Survey of clustering algorithms, “Neural Networks, IEEE Transactions on, vol , 16 .pp. 645_678, 2005.
[14] J. Handl and B. Meyer, "Ant-based and swarm-based clustering," Swarm Intelligence, vol. 1, pp. 95-113, 2007/12/01 2007.
[15] E. R. Hruschka, et al., "Evolving clusters in gene-expression data, " Information Sciences, vol. 176 ,pp. 1898-1927, 2006.
[16] P. S. Shelokar, et al., "An ant colony approach for clustering," Analytica Chimica Acta, vol. 509, pp. 187-195, 2004.
[17] D. W. van der Merwe and A. P. Engelbrecht, "Data clustering using particle swarm optimization," in Evolutionary Computation, 2003. CEC '03. The 2003 Congress on, 2003, pp. 215-220 Vol.1.
[18] P-N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Pearson Addison-Wesley, 2006.
[19] TheLife o f Birds,Parenthood. http://www.pbs.org/lifeofbirds/home/index.html (retrieved 05.09.09