دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)
دانشکده ریاضی و علوم کامپیوتر

پروژه

علوم کامپیوتر

بررسی روش خوشه بندی در مقاله Swarm Intelligence for Self-Organized Clustering

نگارش

مهدی عباسعلی پور

استاد راهنما

جناب آقای دکتر قطعی

استاد مشاور

جناب آقای یوسفی مهر

دی ماه ۱۴۰۲


بسم الله الرحمن الرحيم

چکیده

در این پروژه هدف بر بررسی یک روش مبتنی بر هوش جمعی برای خوشه بندی دیتا می باشد . این روش در مقاله ی Swarm Intelligence for Self-Organized Clustering توضیح داده شده است . روش Databionic که ترکیبی از هوش جمع یهمراه با نظریه بازی ها می باشد . روش مقاله ی مذکور به زبان R پیاده سازی شده است و در این گزارش بخش هایی از روش نیز با زبان پایتون پیاده سازی شده است .

واژه‌های کلیدی:

هوش جمعی ، خودسازمانی ،خوشه بندی ، Databionic

فهرست مطالب

عنوان صفحه

فهرست تصاویر

شکل صفحه

فصل اول

معرفی روش

۱-۱ مقدمه

بسیاری از پیشرفت‌های فناوری با کمک الگوبرداری از طبیعت به صورت استفاده از روش‌ها و سیستم‌های بیولوژیکی موجود در طبیعت ظاهر می‌شود ، پدید آمده اند [۲] . الگویتم های متفاوت هوش مصنوعی همانند الگوریتم های ژنتیک و کلونی مورچگان با الهام از طبیعت مورد استفاده قرار گرفته اند . این موضوع به طور قابل توجهی در مورد الگوریتم های هوش جمعی صدق می کند .به یکی از این دسته الگوریتم ها در [۲] پرداخته شده است . هدف گزارش بر بررسی این روش و همین طور پیاده سازی بخش هایی از این الگوریتم با زبان پایتون و اعمال آن برای روی دیتاست Iris که داده های مربوط به گل هاست ، می باشد . این داده شامل ۱۵۰ نمونه می باشد .

شکل ۱-۱: الگوریتم کلونی مورچگان به عنوان یک الگوریتم هوش جمعی برگرفته شده از طبیعت [۱]

۱-۲ طرح مسئله

الگوریتم مورد بحث در [۲] الگوریتم Databionic می باشد . هدف این الگوریتم خوشه بندی دیتا به صورت بدون نظارت می باشد با استفاده از روش های هوش جمعی می باشد . این الگوریتم در دسته ی الگوریتم های ACS یعنی الگوریتم هایی که عامل های هوشمند به صرت غیر مستقیم با هم در ارتباط اند . قرار دارد .همچنین این الگوریتم ABC [۱] می باشد . این الگوریتم به وسیله ی پارامتر های کمی داده ها را خوشه بندی می نماید .

[۱] ant-based clustering

۱-۳ مراحل الگوریتم خوشه بندی

این الگوریتم از سه مرحله ی ۱- تصویر کردن داده ۲- ساخت Generalized U-matrix ۳- خوشه بندی تقسیم می شود .در ادامه به این مراحل خواهیم پرداخت .

۱-۳-۱ Pswarm

گام اول برای اجرای خوشه بندی تصویر کردن داده ها به فضای دوبعدی می باشد . Polar swarm این تصویر کردن را ممکن می سازد . Pswarm بر پایه ی سه موضوع هوش جمعی ، خود سازمان دهی(به معنای آن که تابع هدفی برای بهینه سازی در نظر گرفته نمی شود و خود سیستم به صورت خودمختار خودش را سازمان دهی می کند) و نظریه تعادل نش در بازی های بدون همکاری استوار است . این الگوریتم بر این مبنا عمل می کند که ساختار داده ی دارای ابعاد بالا را بر رویفضای تصویر شده حفظ نماید به این صورت که فواصل دو به دوی بین دیتای تصویر شده مشابه داده ورودی بماند .

ابتدا بر روی فضای خروجی تعدادی عامل هوشمند آزاد می کند و سپس این عوامل طبق سازوکار هوش جمعی حرکت می کنند و سعی در ساختن ساختار مشابه فضای ورودی دارند . در هر گام تعدادی از عامل ها با احتمالی انتخاب می شوند (این احتمال در گام های بعدی به تدریج کاهش می یابد تا دیگر هیچ عاملی انتخاب نشود) و سپس با توجه به تابع رایحه [۲] انتخاب می کند که به کدام موقعیت برود و فرایند بار ها انجام می شود . طبق نظریه ی نش این عمل همانند یک بازی غیرهمکارانه می باشد و تغییر وضعیت هر عامل بر روی تابع رایحه اثر می گذارد و طبق این نظریه پس از چندین گام همگرایی رخ می هد و سیستم به تعادل می رسد یعنی این همگرایی طبق نظریه نش تضمین شده است . تابع سودمندی که برای بازی در نظر گرفته می شود در حقیقت همان تابع رایحه است که در مقاله به آن اشاره شده است و از آن با λ نام برده شده است .

از مشکلات مهم موجود در فضا به مشکلاتی که می توانند بر روی مرز ناحیه خروجی رخ دهد می توان اشاره کرد . فضای مسئله را به صورت متقارن و در مختصات قطبی در نظر گرفته می شود و همین طور شبکه را به صورت تناوبی تکرار می شود تا مشکلات مرزی رخ ندهد .

ابعاد فضای خروجی

در [۲] برای بدست آوردن ابعاد فضا با توجه به فضای ورودی موردی را ذکر می نماید . سه شرط مهم برای ابعاد فضا وجود دارد . شرط اول مربوط امکان پذیر بودن تبدیل فواصل دیتای ورودی به خروجی می باشد. طوری در نظر گرفته می شود که بیشترین فاصله در فضای خروجی بزگتر یا مساوی نسبت بیشترین فاصله ها به کمترین فاصله ها(به نوعی رزولوشن تقسیم بندی) باشد و این بیشترین فاصله ها و کمترین فاصله ها با صدک ۹۹ ام $p_{99}(\tilde{D})$ و صدک اول $p_{01}(\tilde{D})$ در ماتریس فواصل ورودی تعیین می شود . در ۱-۱ L, C به ترتیب سطر ها وستون های شبکه ی خروجی اند .

[۲] scent

$$\sqrt{C^2 + L^2} \geq \frac{p_{99}(\tilde{D})}{p_{01}(\tilde{D})} =: A \qquad (۱-۱)$$

همچنین ابعاد فضا باید طوری باشد تا عامل ها(دیتابات ها) بتوانند به طور مناسب جهش کنند . این شرط را به صورت زیر بیان می کنیم .

$$L \cdot C \geq \alpha \cdot N \qquad (۱-۲)$$

که در ۱-۲ Nتعداد دیتا بات ها و αتعداد مکان ها برای جهش می باشد . و د آخر شرط سومی به صورت زیر :

$$\frac{1}{L}\frac{C}{\beta} = 1 \qquad (۱-۳)$$

که β مقیاس [۳] شرط سوم برای تضمین مستطیلی بودن شبکه است تا مربعی بودن زیرا نویسنده بر این باور است که فضای مسطیلی عملکرد بهتر ی تا مربعی دارد.
این شرط ها منجر به تعیین علامت معادله سهمی زیر می شود .

$$4C^2 - A^2C^2 + \alpha^2N^2 = 0 \qquad (۱-۴)$$

و در نهایت تعداد ستون های شبکه به صورت ۱-۵ حاصل می شود :

$$C = \begin{cases} \frac{1}{\sqrt{2}}\sqrt{A^2 + \sqrt{A^4 - \frac{\alpha^2}{4}N^2}}, & \square\square A^4 > \frac{\alpha^2}{4}N^2 \\ \text{approximation, otherwise} \end{cases} \qquad (۱-۵)$$

۱-۳-۲ Generalized U-matrix

تا به اینجا تصویر کردن دیتا به دو بعد ساختار فضا را حفظ نموده است . برای تهیه ی تصویر توپوگرافی [۴] نیاز به تطبیق چگالی فضای ورودی با فضای دوبعدی خروجی داریم . این امر به وسیله ی Generalized U-matrix محقق می شود .

[۳] scale
[۴] Topographic map

۱-۳-۳ Clustering

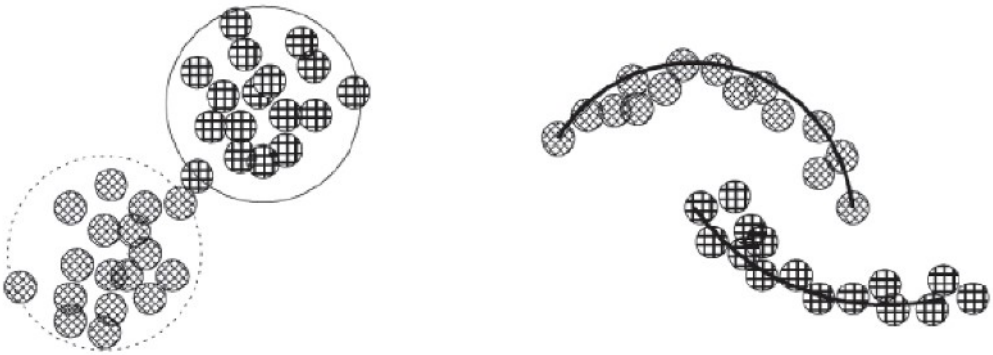
در نهایت می توان با یک نقشه توپوگرافی ساخت که به وسیله ی آن می توان دیدی کلی از داده ها بدست آورد . به وسیله ی این دید می توان خوشه های داده ها را مشاهده نمود . در حقیقت در این بخش می توان با استفاده از توانایی درک انسان الگوهای موجود در دیتا را یافت .

۱-۴ لینک گیت هاب کد

با مراجعه به https://github.com/mahdialipoo/AI_project7 می توانید کد مربوط به پیاده سازی پروژه را مشاهده نمایید .

۱-۵ جمع بندی

در این فصل به بررسی مراحل الگوریتم خوشه بندی Databionic پرداخته شد .این روش یک روش کم پارامتر می باشد که تنها با دو پارامتر تعداد خوشه ها و نوع خوش بندی (پیوسته یا فشرده ۱-۲) اقدام به خوشه بندی داده ها می نماید . روش Databionic یک روش نیمه تعاملی می باشد که با تهیه نقشه توپوگرافی کمک می کند تا بتوان پارامتر های مورد نیاز برای خوشه بندی را تعیین نمود .



شکل ۱-۲: دو نع خوشه بندی فشرده (سمت چپ) و خوشه بندی هبند (سمت راست) [۳]

فصل دوم

پیاده سازی بخش Pswarm و بررسی نتایج

۲-۱ پیاده سازی

به علت پیچیدگی بخش های دیگر الگوریتم تنها به پیاده سازی قسمت ابتدایی یعنی تصویر کردن داده ها به دو بعد اکتفا می کنم .

۲-۱-۱ Databot Class

برای پیاده سازی در ابتدا کلاس Databot تعریف شد . هر آبجکت دیتا بات یک موقعیت در صفحه دوبعدی دارد و مین طور با توجه به موقعیتش در متغیر payoff میزان رایحه scent را ذخیره می نماید . همچنین هر آبجکت یک تابع دارد تا با استفاده ا زموقعیت فعلی آبجک دیتا بات و بادر نظرگرفتن موقعیت بقیه ی دیتابات ها با استفاده از H-weight تابع سودمندی را بروز رسانی می نماید (متد update-payoff)

۲-۱-۲ تبدیل مختصات

تابعی برای تبدیل مختصات قطبی به دکارتی به نام distance-2D-polar در نظرگرفته شده است که مختصات را به دکارتی می برد .

۲-۱-۳ برروز رسانی موقعیت

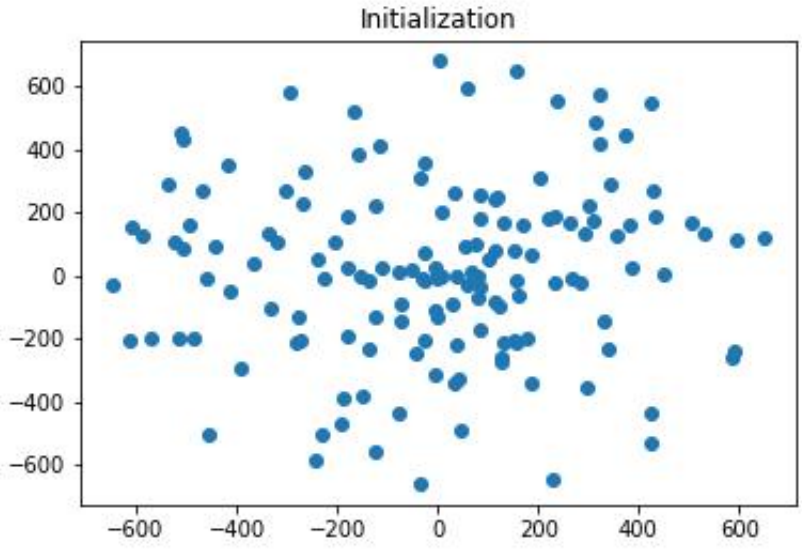
تابع update-positions مطابق روش مطرح شده به صورت تبریدی نمونه ای از دیتابات ها را انتخاب می کند و در صورتی که بتوانند به موقعیتی با تابع سودمندی بهتری بروند موقعیت آن ها را تغییر می دهد .

۲-۱-۴ تابع Pswarm

این تابع به عملیات اصلی تصویر کردن می پردازد و تکرار های حل مسئله را انجام می دهد . به صورت تبریدی شروع به فراخوانی توابع بروزرسانی موقعیت می نماید و داده ها را تصویر می کند . در ۲-۱ و ۲-۲ تغییر مکان دیتا بات ها را مشاهده می نمایید .

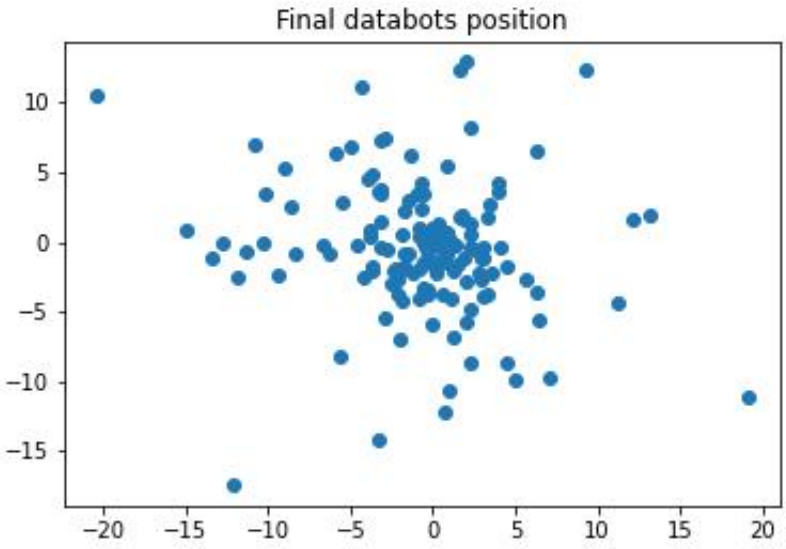
۲-۲ جمع بندی

در این بخش بررسی برخی از نتایج پیاده سازی بخش pswarm از الگوریتم خوشه بندی Databionic پرداختیم . همان طور که در ۲-۱ و ۲-۲ مشاهده می شود داده ها به سمت یک نقطه جمع شده اند مجموع تابع سودمندی دیتا بات ها از مقدار اولیه ی ۱۵.۳۷۹ به مقدار ۳.۳۷۹ رسید .مدت زمان اجرا شدن الگوریتم تنها برای ۱۵۰ داده در حدود ۲دقیقه و ۴۰ ثانیه طول کشید که نشان می دهد بسیار کند عمل شده است . همین طور به نظر نمی رسد که تصویر کردن به این صورت با فرض ساده سازی



شکل ۲-۱: پراکندگی اولیه ی عامل ها(دیتا بات ها)

های انجام شده (مانند متناوب نگرفتن فضای خروجی) و در صورت ادامه ی باقی بخش های الگوریتم به نتایج مورد قبولی ختم شود .



شکل ۲-۲: پراکندگی نهایی عامل ها(دیتا بات ها)

مراجع

[1] sda apa. Researchers develop robotic ants with swarm intelligence. , 2019.

[2] Thrun, Michael C and Ultsch, Alfred. Swarm intelligence for self-organized clustering. *Artificial Intelligence*, 290:103237, 2021.

[3] Thrun, Michael Christoph. *Projection-based clustering through self-organization and swarm intelligence: combining cluster analysis with the visualization of high-dimensional data*. Springer, 2018.