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

پروژه

علوم کامپیوتر

خوشه بندی دیتاست Iris با استفاده از الگوریتم ژنتیک و مقایسه ی آن با Kmeans

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 این است تا دیتاست گل های Iris که مربوط به گل ها می باشند را با استفاده از روش الگوریتم ژنتیک به صورت تکاملی خوشه بندی کنم و آن را با الگوریتم متداول Kmeans مقایسه کنم .

واژه‌های کلیدی:

الگوریتم ژنتیک،خوشه بندی، Kmeans ، الگوریتم ها تکاملی

فهرست مطالب

عنوان صفحه

فهرست تصاویر

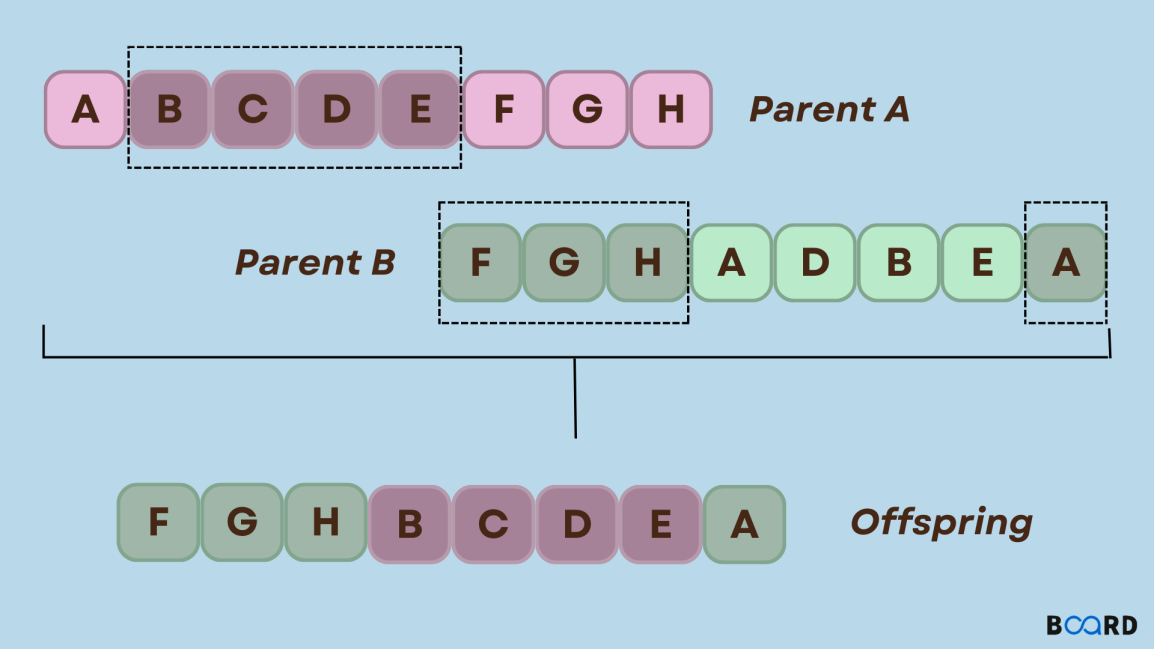
شکل صفحه

فصل اول

معرفی مسئله

۱-۱ مقدمه

امروزه مقالات متعددی [۲] و [۳] در باره ی استفاده از الگوریتم ژنتیک برای بهینه سازی توابع مطلوبیت در الگوریتم های یادگیری ماشین و دسته بندی دیتا ست ها نوشته می شود . در این پروژه هدف بر این می باشد تا در مورد دیتا ست گل های Iris که توسط کتابخانه ی sklean قابل دسترسی می باشد عملکرد دو الگوریتم تکاملی و Kmeans را مقایسه نماییم .



شکل ۱-۱: ترکیب در الگوریتم ژنتیک
[۱]

۱-۲ داده ها

مجموعه داده گل زنبق (به انگلیسی: Iris flower data set) یا مجموعه داده زنبق فیشر یک مجموعه داده چند متغیره است که توسط رانلد فیشر، آماردان و زیست‌شناس بریتانیایی در سال ۱۹۳۶ معرفی شد. این مجموعه داده همچنین مجموعه داده زنبق اندرسون نیز نامیده می‌شود. این مجموعه شامل ۱۵۰ نمونه‌ی جمع‌آوری شده از گل‌های زنبق است که این نمونه‌ها ۵۰ نمونه از هر یک از سه نوع گل زنبق را شامل می‌شوند. برای هر یک از نمونه‌ها ۴ ویژگی گل زنبق اندازه‌گیری شده‌است. این ویژگی‌ها شامل طول و عرض کاسبرگ و گلبرگ، بر حسب سانتی متر است [۴] .

۱-۳ لینک گیت هاب کد

با مراجعه به https://github.com/mahdialipoo/AI_project6 می توانید کد مربوط به پیاده سازی پروژه را مشاهده نمایید .

فصل دوم

پیاده سازی

در این بخش به توابعی که برای استفاده از الگوریتم ژنتیک استفاده کردم می پردازم . نا گفته نماند که در نوشتن برخی از این توابع از چت جی پی تی استفاده شده است . در ابتدا ی امر دیتای مورد نظر را از sklearn بارگیری می نماییم .

۲-۱ مقدار دهی اولیه

مقدار دهی اولیه جمعیت به صورت تصادفی انجام می شود . هر کروموزوم مطابق آنچه در راهنما گفته شده است پر می شوند . جمعیت مورد نظر را ۵۰ در نظر می گیریم و سپس به اندازه ۱۰۰ نسل فرایند را تکرار می نماییم .

۲-۲ تابع سازگاری

تابع ساز گاری ، معیار نسبت واریانس Calinski-Harabasz قرار داه شده است . که مجموع نسبت پراکندگی داده های داخل هرخوشه را نسبت به پراکندگی به داده های خوشه های دیگر می سنجد .

۲-۳ تابع ترکیب

این تابع را مطابق با حالت استانداردی که داخل اسلاید های درس گفته شده است نوشته شده است . دو کروموزوم توسط تابع انتخاب ، انتخاب می شوند و تعدادی از ژن هایشان را جابجا می کنیم و فرزندان آن ها ایجاد می شوند. از بین خودشان و فرزندانشان با توجه به تابع سازگاری(این تابع توزیع احتمال انتخاب را تعیین می کند) دو کروموزوم را به جمعیت اضافه می کنیم .

۲-۴ تابع انتخاب

این تابع دو کروموزوم را به صورت احتمالاتی انتخاب می کند و هرچه مقدار تابع ساز گاری برای کروموزومی بیشتر باشد احتمال آن که انتخاب شود نیز بیشتر می باشد .

۲-۵ تابع جهش

در تابع جهش هم به صورت تصادفی تعدادی از جمعیت را انتخاب می کنیم و خوشه ی یکی از دیتا پوینت ها را با مقدار تصادفی عوض می کنیم .

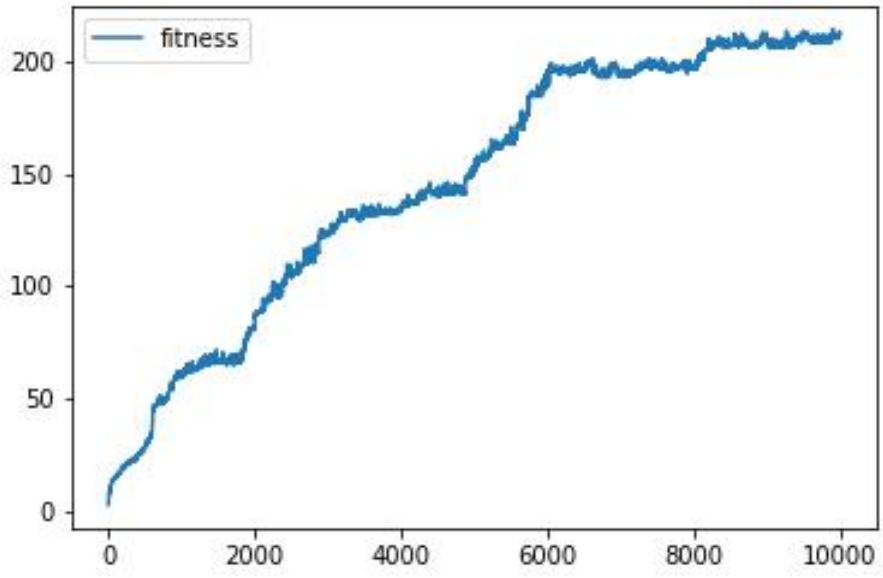
۲-۶ جمع بندی

در بخش فعلی بخش های متفاوت کدپیاده سازی شده را توضیح دادیم . این بخش ها (تابع جهش و ترکیب و انتخاب و...)بر عمما بر اساس شیوه ی متداول برای الگوریتم ژنتیک و اسلاید های درس پیاده شده اند .

فصل سوم

نتایج و ارزیابی

در ۳-۱ تغییرات تابع سازگاری برای بهترین نمونه ی جمعیت مشاهده می کنید. نوساناتی وجود دارند . هم گرایی بسار کند صورت گرفت . زمان انجام الگوریتم برای ۱۰۰۰۰ نسل و جمعیت۱۵۰ در حدود ۱ساعت زمان برد .

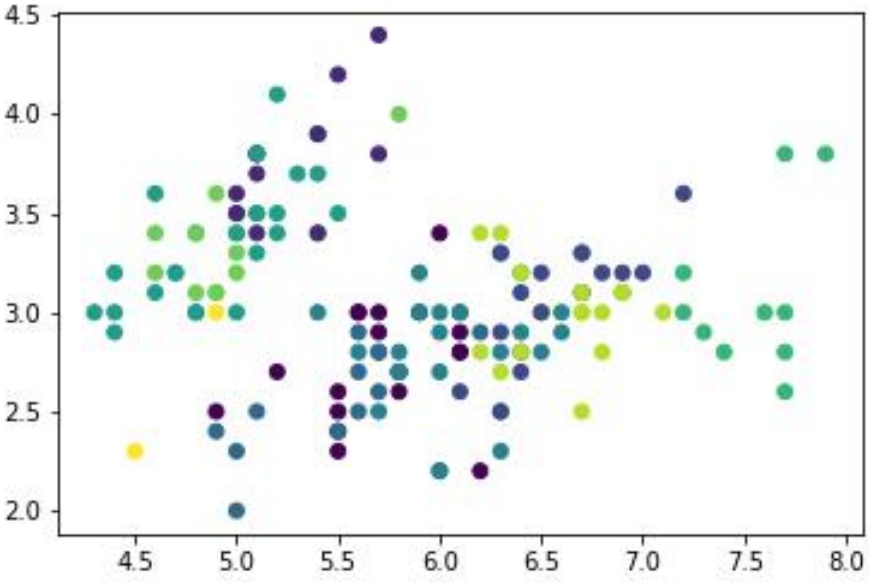


شکل ۳-۱: تغییرات مقدار بیشینه سازگاری جمعیت در طول نسل های مختلف

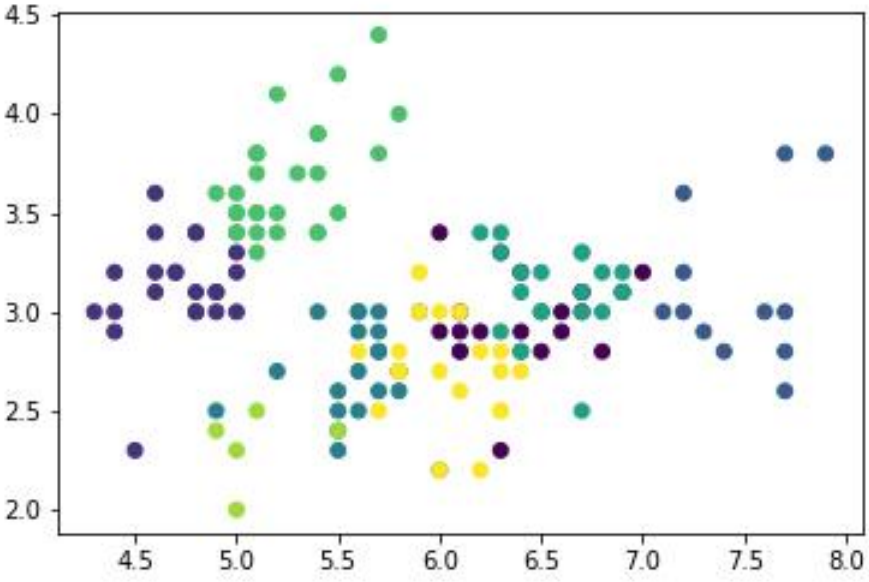
در ۳-۲ و ۳-۳ خوشه بندی دیتا پوینت ها را برای داده ها به ترتیب به روش های تکاملی و Kmeans مشاهده می کنید. در ۳-۴ نیز خوشه ها بر اساس داده های دیتا ست نمایش داده شده است .

۳-۱ مقایسه ی مدل ها

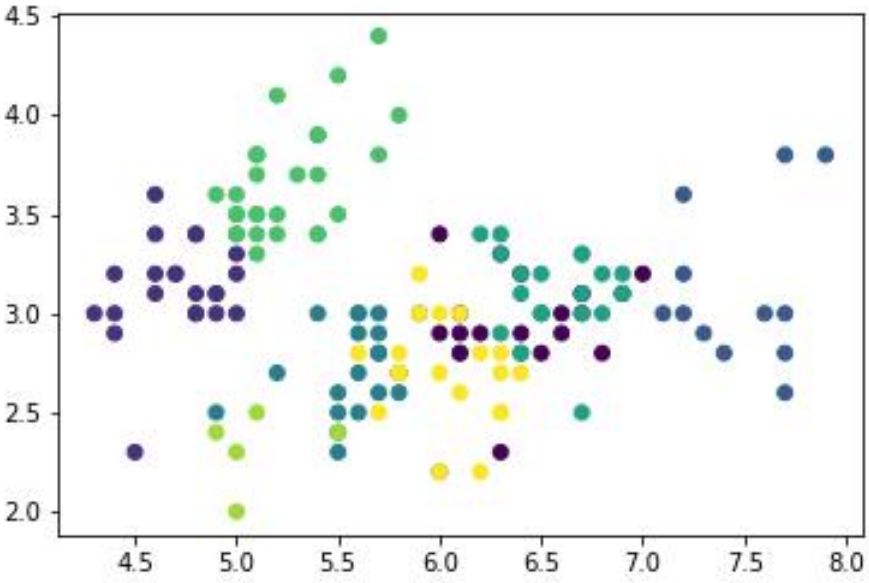
زمان ساخته شدن مدل به صورت تکاملی در مقایسه با Kmeans بیشتر طول کشد . مقدار تابع سازگاری در نهایت برای روش تکاملی مقدار 214 شد و برای Kmeans مقدار آن 440 شد . این نتایج نشان دهنده عملکرد بهتر Kmeans نسبت به حالت تکاملی بر روی داه های این دیتاست می باشد .



شکل ۳-۲: خوشه بندی بر اساس دو ویژگی دلخواه بر اساس روش تکاملی



شکل ۳-۳: خوشه بندی بر اساس دو ویژگی دلخواه بر اساس روش Kmeans



شکل ۳-۴: خوشه ها بر اساس خود دیتاست

مراجع

[1] boardinfinity. Genetic algorithm in machine learning. https://www.boardinfinity.com/blog/genetic-algorithm-in-machine-learning/, 2023.

[2] Hruschka, Eduardo R and Ebecken, Nelson FF. A genetic algorithm for cluster analysis. *Intelligent data analysis*, 7(1):15–25, 2003.

[3] Waboke, William Rupert, Bagiwa, Mustapha Aminu, Obiniyi, Ayodele Afoloyan, and Obasa, Adekunle Isiaka. Centroid initialization in k-means clustering using gatcam. *Science World Journal*, 18(1):143–151, 2023.

[4] wikipedia. Iris flower data set. https://en.wikipedia.org/wiki/Iris-flower-data-set, 1936.