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

پروژه درس هوش مصنوعی و کارگاه

خوشه‌بندی با استفاده از الگوریتم ژنتیک

نگارش
ایلیا راوند

استاد درس
دکتر مهدی قطعی

استاد کارگاه
بهنام یوسفی مهر

بهار ۱۴۰۳


بسم الله الرحمن الرحيم

چکیده

در این پروژه ما به بررسی نحوه پیاده‌سازی الگوریتم ژنتیک می‌پردازیم، در ادامه استفاده آن در مسئله‌های خوشه‌بندی را مشاهده می‌کنیم و با الگوریتم K-Means مقایسه می‌کنیم.

واژه‌های کلیدی:

خوشه‌بندی، الگوریتم ژنتیک و K-Means

فهرست مطالب

عنوان صفحه

فهرست اشکال

شکل صفحه

فصل اول

مقدمه

مسئله ما پایگاه‌داده‌ای را در نظرگرفته‌است که شامل ۷ نوع داده می‌باشد. می‌خواهیم استفاده الگوریتم ژنتیک را در مسائل خوشه‌بندی بررسی کنیم و پیاده‌سازی مدلی از آن را ببینیم. پس در ابتدا توضیح مختصری از نحوه پیاده‌سازی الگوریتم ژنتیک می‌دهیم و نحوه ارتباط آن را به این نوع مسائل بررسی می‌کنیم. بعد به پیاده‌سازی آن می‌پردازیم و در انتها به مقایسه آن با K-Means می‌پردازیم.

فصل دوم

الگوریتم ژنتیک

الگوریتم ژنتیک روشی برای پیدا کردن بهترین راه‌حل با الگو گرفتن از طبیعت است. مراحل اصلی الگوریتم ژنتیک به طور خلاصه به صورت زیر تعریف می‌شود:

تعریف کردن مفهوم یک راه‌حل به صورت ترجیحاً عددی که به آن کروموزوم می‌گوییم. هر کروموزوم شامل چندین ژن است که ژن در واقع بخش‌های خرد شده یک راه حل هستند. در ابتدا ما چندین کروموزوم می‌سازیم که به آن جمعیت اولیه می‌گویند. بعد با استفاده از یک تابع هریستیک، چند نفر از این جمعیت را به عنوان والد انتخاب می‌کنیم که به آن انتخاب می‌گویند. حال از افراد انتخاب شده به طور شانسی دو نفر را انتخاب کرده و آن دو کراس‌اور [1] انجام می‌دهند به معنای این که فرزند تولید می‌کنند. این فرزند باید از روی ژن‌های والد خود ساخته شود و در نهایت یک پَرِش ژن که در تابع موتاسیون [2] اتفاق می‌افتد، انجام بپذیرد. این کار را انقدر تکرار می‌کنیم که جمعیت اولیه دوباره بازسازی شود.

[1] Crossover

[2] Mutation

فصل سوم

تحلیل رابطه‌ی خوشه‌بندی و الگوریتم ژنتیک

اکنون سوال مطرح شده این می‌باشد که ارتباط الگوریتم ژنتیک به مسئله خوشه‌بندی چیست؟ در واقع با اعمال یک بخش روی هر دیتابیس می‌توان سوال خوشه‌بندی را با الگوریتم ژنتیک حل کرد و به آن بخش انکودینگ می‌گویند. یعنی ما مشخص کنیم که در مسئله خوشه‌بندی منظورمان از راه حل چیست؟ چه چیزی برای ما در ارزیابی بهترین راه حل مفهوم دارد؟ و چه داده‌هایی داریم؟

با پاسخ دادن به این سوال‌ها ما می‌توانیم یک انکودینگ مناسب برای الگوریتم ژنتیک درست کنیم. به طور خلاصه مراحل اصلی به این صورت است:

۱. تعریف کروموزوم

هر کروموزوم نمایانگر یک راه حل ممکن است. در مسئله خوشه‌بندی، هر کروموزوم می‌تواند نمایانگر تخصیص داده‌ها به خوشه‌های مختلف باشد. برای مثال، هر ژن در کروموزوم می‌تواند یک عدد باشد که نمایانگر خوشه‌ای است که داده به آن تخصیص داده شده است.

۲. جمعیت اولیه

مجموعه‌ای از کروموزوم‌ها (راه‌حل‌ها) را به صورت تصادفی یا با استفاده از روش‌های خاصی ایجاد می‌کنیم.

۳. تابع برازش (Fitness Function)

یک تابع که کیفیت هر کروموزوم (راه حل) را ارزیابی می‌کند. در مسئله خوشه‌بندی، این تابع می‌تواند بر اساس فاصله درون خوشه‌ای و فاصله بین خوشه‌ای باشد.

۴. انتخاب

تعدادی از بهترین کروموزوم‌ها را برای تولید نسل بعد انتخاب می‌کنیم. این انتخاب می‌تواند بر اساس روش‌های مختلفی مانند تورنمنت، رتبه‌بندی یا روش‌های دیگر باشد.

۵. تولید نسل جدید

با استفاده از عملگرهای ژنتیکی مانند کراس‌اور (ترکیب کروموزوم‌های والد برای ایجاد کروموزوم‌های فرزند) و موتاسیون (ایجاد تغییرات تصادفی در کروموزوم‌ها) نسل جدیدی از کروموزوم‌ها را ایجاد می‌کنیم. این فرایند شامل مراحل زیر است:

- کراس‌اور

دو کروموزوم والد انتخاب می‌شوند و بخشی از ژن‌های آن‌ها با هم ترکیب می‌شوند تا یک یا دو کروموزوم فرزند تولید شود. این ترکیب باید به گونه‌ای باشد که ویژگی‌های خوب والدین را به فرزندان منتقل کند.

- موتاسیون

برخی از ژن‌های کروموزوم فرزند به صورت تصادفی تغییر می‌کنند تا تنوع ژنتیکی در جمعیت حفظ شود و احتمال یافتن راه‌حل‌های بهتر افزایش یابد.

۶. ارزیابی و تکرار

نسل جدید کروموزوم‌ها با استفاده از تابع برازش ارزیابی می‌شوند و این فرایند تکرار می‌شود تا زمانی که یکی از معیارهای توقف (مثلاً تعداد نسل‌های مشخص شده یا رسیدن به یک حداقل بهبود در تابع برازش) برآورده شود.

با این روش، الگوریتم ژنتیک می‌تواند راه‌حل‌های بهینه یا نزدیک به بهینه را برای مسئله خوشه‌بندی پیدا کند. در نهایت، بهترین کروموزوم (راه حل) که دارای بالاترین مقدار برازش است به عنوان بهترین تخصیص داده‌ها به خوشه‌ها انتخاب می‌شود.

تعریف کروموزوم که مهم‌ترین بخش است، در واقع رابطه بین الگوریتم ژنتیک و مسئله خوشه‌بندی است. در این مسئله هر کروموزوم شامل ۷ ژن است که هر ژن مختصات مرکز یک خوشه است و بقیه الگوریتم بر روی این ساختار ساخته می‌شود.

فصل چهارم

تحلیل رابطه‌ی **K-Means** و الگوریتم ژنتیک

تفاوت اصلی بین K-Means و پیاده‌سازی الگوریتم ژنتیک برای خوشه‌بندی در نحوه کارکرد و روش‌های بهینه‌سازی آن‌هاست. در ادامه تفاوت‌های کلیدی این دو روش را بیان می‌کنیم:

۱. روش بهینه‌سازی

- K-Means

یک الگوریتم قطعی [۱] است که بر پایه تکرار و به‌روزرسانی مراکز خوشه‌ها [۲] عمل می‌کند. هدف آن کمینه‌کردن مجموع مربعات فواصل داده‌ها تا نزدیک‌ترین مرکز خوشه است. این الگوریتم با یک مجموعه اولیه از مراکز خوشه شروع می‌شود و در هر مرحله داده‌ها را به نزدیک‌ترین مرکز خوشه اختصاص می‌دهد و سپس مراکز خوشه را بر اساس میانگین داده‌های تخصیص داده شده به‌روزرسانی می‌کند. این فرآیند تا همگرا شدن مراکز خوشه‌ها ادامه می‌یابد.

- Genetic Algorithm

یک الگوریتم تصادفی و الهام گرفته از طبیعت است که با استفاده از اصول انتخاب طبیعی، کراس‌اور و موتاسیون بهینه‌سازی می‌کند. الگوریتم ژنتیک با یک جمعیت اولیه از راه‌حل‌ها شروع می‌شود و از طریق تولید نسل‌های جدید، سعی می‌کند راه‌حل‌های بهینه‌تری پیدا کند. هر نسل شامل اعمال انتخاب، کراس‌اور و موتاسیون است که به تدریج جمعیت را به سمت راه‌حل‌های بهتر هدایت می‌کند.

۲. همگرایی و سرعت

- K-Means

معمولاً سریع‌تر از الگوریتم ژنتیک است و در بسیاری از موارد به سرعت همگرا می‌شود. اما ممکن است به یک نقطه بهینه محلی برسد و بهینه کلی را پیدا نکند. نتایج K-Means به شدت به مراکز اولیه خوشه‌ها وابسته است.

- Genetic Algorithm

نسبت به K-Means کندتر است، زیرا نیاز به انجام چندین نسل دارد تا به یک راه‌حل بهینه برسد. با این حال، به دلیل طبیعت تصادفی و جستجوی گسترده‌تر، الگوریتم ژنتیک معمولاً شانس بیشتری برای پیدا کردن بهینه جهانی دارد و کمتر در بهینه‌های محلی گیر می‌افتد.

[1] Deterministic

[2] Centroids

۳. پیچیدگی

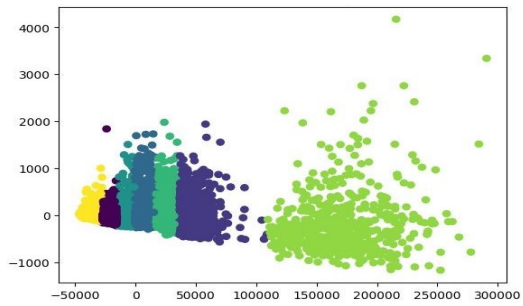
- K-Means

نسبتاً ساده‌تر و سرراست‌تر است. تنها نیاز به تعریف مراکز خوشه‌ها و تخصیص داده‌ها به خوشه‌ها دارد.

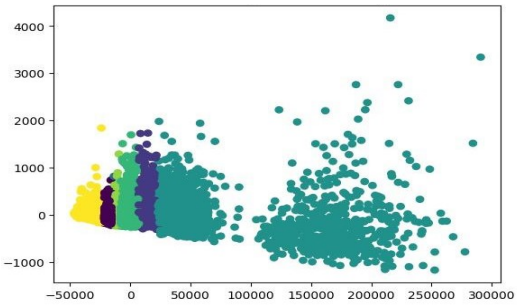
- Genetic Algorithm

پیچیده‌تر است و نیاز به تعریف دقیق کروموزوم‌ها، تابع برازش، و عملگرهای ژنتیکی (کراس‌اور و موتاسیون) دارد. همچنین پارامترهای مختلفی مانند اندازه جمعیت، نرخ موتاسیون و نرخ کراس‌اور باید تنظیم شوند.

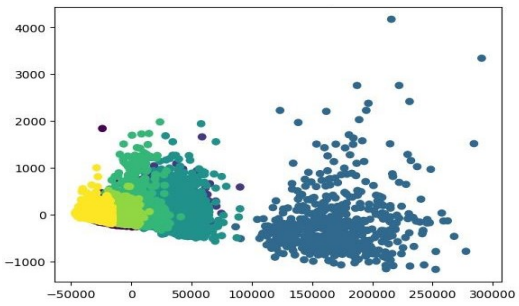
به طور کلی، K-Means یک روش ساده و کارآمد برای مسائل خوشه‌بندی استاندارد است، در حالی که الگوریتم ژنتیک انعطاف‌پذیری بیشتری دارد و می‌تواند مسائل پیچیده‌تر و بهینه‌سازی‌های چند هدفه را بهتر مدیریت کند. خروجی حاصل شده در شکل‌های ۴-۱، ۴-۲ و ۴-۳ قابل مشاهده است که در آن میزان دقت K-Means و الگوریتم ژنتیک به ترتیب برابر با ۳٪ و ۵٪ می‌باشد.



شکل ۴-۱: K-Means Clustering



شکل ۴-۲: GA Clustering



شکل ۴-۳: True Clustering

فصل پنجم

پیاده‌سازی توابع الگوریتم ژنتیک

در این فصل به معرفی توابع مورد نیاز برای پیاده‌سازی الگوریتم ژنتیک می‌پردازیم.

- initialize_population

با گرفتن کل داده‌ها و طول هر کروموزوم و تعداد کروموزوم‌ها، به ما یک جمعیت را برمی‌گرداند که به صورت تصادفی انتخاب شده‌اند.

- fitness_function

هر کروموزوم را به نزدیک‌ترین مرکز نسبت می‌دهد و بعد راه حلی که کروموزوم به آن اشاره دارد را با جواب اصلی مسئله که در پایگاه داده بوده مقایسه می‌کند.

- select_parents

همه را بر اساس fitness مرتب می‌کند و به تعداد خواسته شده از بهترین‌ها برمی‌دارد.

- crossover

دو والد را گرفته و یک نقطه به طور شانسی بین ژن‌ها انتخاب می‌کند و بچه تولید شده از ابتدا تا آن ژن خاص یک والد و از آنجا تا انتهای والد دیگر را به ارث می‌برد.

- mutate

با گرفتن یک فرزند، هر ژنش را به احتمال مشخصی یک عدد تصادفی می‌گذارد.