دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده ریاضی و علوم کامپیوتر

پروژه پنجم هوش مصنوعی

رشته علوم کامپیوتر

تحلیل دیتا

نگارش

علیرضا مختاری

استاد درس

مهدی قطعی

استاد کارگاه

بهنام یوسفی مهر

مهر ۱۴۰۳

چکیده

در این پروژه، از الگوریتم ژنتیک برای حل مسأله خوشه‌بندی داده‌ها و مقایسه آن با الگوریتم k-Means استفاده شده است. هدف اصلی، ارزیابی عملکرد الگوریتم ژنتیک در خوشه‌بندی داده‌ها و بررسی کیفیت نتایج به کمک معیار سیلوئت (Silhouette Score) بوده است. برای این منظور، یک دیتاست واقعی پیش‌پردازش و نرمال‌سازی شد و سپس الگوریتم ژنتیک برای تعیین مراکز خوشه‌ها به کار گرفته شد. نتایج حاصل از الگوریتم ژنتیک با خروجی k-Meansمقایسه گردید. ارزیابی‌ها نشان داد که الگوریتم ژنتیک در شناسایی خوشه‌ها به صورت مؤثر عمل می‌کند و توانایی رقابت با k-Meansرا دارد. این تحقیق نشان‌دهنده پتانسیل الگوریتم‌های تکاملی برای حل مسأله‌های پیچیده خوشه‌بندی است.

واژه‌های کلیدی:

الگوریتم ژنتیک ، خوشه‌بندی ، k، k-Means ، معیار سیلوئت (Silhouette Score) ، بهینه‌سازی تکاملی

فهرست مطالب

فصل اول

مقدمه

مقدمه

خوشه‌بندی یکی از مسأله‌های مهم در یادگیری بدون نظارت است که در بسیاری از کاربردها، از جمله تحلیل داده‌های آماری، بازاریابی، و بینایی ماشین، مورد استفاده قرار می‌گیرد. الگوریتم k-Means یکی از رایج‌ترین روش‌های خوشه‌بندی است که با استفاده از مراکز اولیه به صورت تصادفی، داده‌ها را به k خوشه تقسیم می‌کند. با این حال، این الگوریتم به شدت به مقادیر اولیه وابسته است و ممکن است به یک مینیمم محلی برسد.

در این پژوهش، از الگوریتم ژنتیک به عنوان یک روش تکاملی برای حل مسأله خوشه‌بندی استفاده شده است. الگوریتم ژنتیک با شبیه‌سازی فرآیند تکامل طبیعی و با استفاده از مفاهیمی همچون انتخاب، ترکیب (Crossover) و جهش(Mutation) ، بهینه‌ترین مراکز خوشه‌ها را پیدا می‌کند. در این پروژه، الگوریتم ژنتیک برای خوشه‌بندی داده‌های یک دیتاست واقعی به کار گرفته شده و نتایج آن با الگوریتم k-Means مقایسه شده است.

.

فصل دوم

مقایسه الگوریتم ها

- الگوریتم **k-means** و خوشه‌بندی با استفاده از الگوریتم‌های تکاملی هر دو روش‌هایی برای خوشه‌بندی داده‌ها هستند، اما تفاوت‌های اساسی در روش کار و کاربردهای آن‌ها دارند.

۲- k-means:

- **روش کار** k-means :یک الگوریتم تکراری است که ابتدا k مرکز به صورت تصادفی یا از پیش تعیین‌شده انتخاب می‌شوند و سپس داده‌ها بر اساس فاصله از این مراکز دسته‌بندی می‌شوند. مراکز به‌روزرسانی می‌شوند تا زمانی که الگوریتم به همگرایی برسد.
- **مزایا** :
  - سرعت بالا و پیچیدگی زمانی کم.
  - پیاده‌سازی ساده و مناسب برای داده‌های کوچک یا متوسط.
- **معایب** :
  - وابستگی به مقدار اولیه.k
  - حساسیت به نویز و داده‌های پرت.
  - در داده‌های غیرخطی عملکرد مناسبی ندارد.
  - همیشه در مینیمم محلی همگرا می‌شود.

۳- خوشه‌بندی با الگوریتم‌های تکاملی:

- **روش کار** :الگوریتم‌های تکاملی مانند الگوریتم ژنتیک (GA) یا بهینه‌سازی ازدحام ذرات (PSO)برای یافتن مراکز خوشه‌ها استفاده می‌شوند. این الگوریتم‌ها به جای یک راه‌حل، جمعیتی از راه‌حل‌ها را در هر مرحله نگهداری کرده و با استفاده از عملیات‌های مانند جهش، ترکیب و انتخاب راه‌حل‌ها را بهبود می‌دهند.
- **مزایا** :
  - مناسب برای داده‌های پیچیده با مرزهای غیرخطی.
  - کاهش وابستگی به مقادیر اولیه.

- توانایی جستجوی جهانی و جلوگیری از گیر افتادن در مینیمم محلی.

- **معایب** :
  - پیچیدگی محاسباتی بالا.
  - زمان اجرای بیشتر نسبت به.k-means
  - نیاز به تنظیم پارامترهای مختلف مانند اندازه جمعیت و نرخ جهش.

٤- مقایسه نهایی:

- اگر داده‌ها ساده و خطی باشند و سرعت اجرا اهمیت داشته باشد، k-meansانتخاب مناسبی است.
- اگر داده‌ها پیچیده، غیرخطی یا دارای نویز باشند، الگوریتم‌های تکاملی عملکرد بهتری خواهند داشت.
- k-meansبرای مسائلی که به محاسبات سریع نیاز دارند، مناسب‌تر است، اما الگوریتم‌های تکاملی انعطاف‌پذیری بیشتری در یافتن خوشه‌های پیچیده‌تر دارند.
- در مجموع، انتخاب بین این دو روش به نوع داده و محدودیت‌های زمانی و محاسباتی بستگی دارد.

فصل سوم

جمع‌بندی و نتیجه‌گیری و پیشنهادات

جمع بندی و نتیجه گیری

در این پروژه، الگوریتم ژنتیک توانست با استفاده از یک رویکرد تکاملی، خوشه‌های داده را به صورت مؤثر تعیین کند. مقایسه معیار سیلوئت بین الگوریتم ژنتیک و k-Means نشان داد که هر دو روش قابلیت مناسبی در خوشه‌بندی داده‌ها دارند، اما الگوریتم ژنتیک به دلیل ماهیت جستجوی سراسری (Global Search)، پتانسیل بیشتری در پیدا کردن خوشه‌های بهینه دارد، به‌ویژه در شرایطی که داده‌ها دارای پیچیدگی بیشتری هستند.

الگوریتم ژنتیک همچنین انعطاف‌پذیری بیشتری در تعریف توابع هدف (Fitness Function) دارد، که می‌تواند بسته به نیاز مسئله، بهبودهای بیشتری را ارائه دهد. با این حال، این روش به دلیل زمان اجرای طولانی‌تر نسبت به k-Means، برای داده‌های بسیار بزرگ نیاز به بهینه‌سازی بیشتری دارد.

در نهایت، این تحقیق نشان داد که استفاده از الگوریتم‌های تکاملی مانند الگوریتم ژنتیک، یک روش قدرتمند و قابل اعتماد برای خوشه‌بندی داده‌ها است که می‌تواند به‌عنوان جایگزینی برای روش‌های سنتی به کار رود.

منابع و مراجع

[1] https://ieeexplore.ieee.org/document/10128470/

[۲] https://link.springer.com/article/XXXXXX

[3] https://scikit-learn.org/

[4] https://github.com/shankarpandala/lazypredict