دانشکده علوم

بخش ریاضی

پایان نامه کارشناسی ارشد

در ریاضی کاربردی

حل مسائل خوشه بندی با استفاده از بهینه سازی

شبیه سازی حرارتی

توسط

زهره السادات ناظمی

استاد راهنما:

دکترکورش زیارتی

استاد مشاور:

دکتر محمد باقراحمدی

دکتر عبدالعزیز عبدالهی

شهریور ۱۳۸۸


بسم الله الرحمن الرحيم

به نام خدا

اظهارنامه

اینجانب زهره السادات ناظمی(۸۵۰۴۲۲) دانشجوی رشته‌ی ریاضی کاربردی گرایش تحقیق در عملیات دانشکده علوم اظهار می‌کنم که این پایان نامه حاصل پژوهش خودم بوده و در جاهایی که از منابع دیگران استفاده کرده‌ام، نشانی دقیق و مشخصات کامل آن را نوشته‌ام. همچنین اظهار می‌کنم که تحقیق و موضوع پایان نامه‌ام تکراری نیست و تعهد می‌نمایم که بدون مجوز دانشگاه دستاوردهای آن را منتشر ننموده و یا در اختیار غیر قرار ندهم. کلیه حقوق این اثر مطابق با آیین نامه مالکیت فکری و معنوی متعلق به دانشگاه شیراز است.

نام ونام خانوادگی: زهره السادات ناظمی

۱۳۸۸/۶/۳۱

به نام خدا

حل مسائل خوشه بندی با استفاده از بهینه سازی شبیه سازی حرارتی

به کوشش

زهره السادات ناظمی

پایان نامه

ارایه شده به تحصیلات تکمیلی دانشگاه شیراز به عنوان بخشی از فعالیت‌های

تحصیلی لازم برای اخذ درجه کارشناسی ارشد

در رشته‌ی

ریاضی کاربردی(تحقیق در عملیات)

از دانشگاه شیراز

شیراز

جمهوری اسلامی ایران

ارزیابی کمیته پایان نامه، با درجه : - - - - - - - -

دکتر کورش زیارتی، استادیار دانشکده مهندسی برق و کامپیوتر - - - - - - - - - -

دکترمحمد باقر احمدی، استادیار داشکده علوم، بخش ریاضی - - - - - - - - - -

دکتر عبدالعزیز عبدالهی، دانشیار دانشکده علوم، بخش ریاضی - - - - - - - - - -

شهریور ۱۳۸۸

تقدیم به :

مادر عزیزم

پدر مهربانم

همسر فداکارم

و خواهران عزیزم

به پاس محبت‌های بی‌دریغشان

سپاسگزاری

با تشکر از استاد محترم جناب آقای دکتر زیارتی که مرا در این امر مهم یاری نمودند و همچنین اساتید گرانقدر جناب آقای دکتر احمدی و دکتر عبدالهی که با راهنمایی های ارزنده خود مسیر را برای من هموار نمودند.

چکیده

حل مسائل خوشه بندی با استفاده از بهینه سازی شبیه سازی حرارتی

به کوشش

زهره السادات ناظمی

خوشه بندی فرایندی است که در طی آن مجموعه‌ای از نمونه‌ها به خوشه هایی تقسیم می‌شوند که اعضای هرخوشه بیشترین شباهت را به یکدیگر داشته باشند و خوشه‌های مختلف با یکدیگر بیشترین تفاوت را داشته باشند. خوشه بندی یکی از تکنیک‌های داده کاوی و آنالیز داده متعارف می‌باشد. درخوشه بندی داده‌ها، در مسائل با اندازه داده بزگتر رسیدن به حل بهینه مشکل‌تر می‌باشد و در نتیجه مدت زمان لازم برای رسیدت به حل‌های قابل قبول طولانی‌تر می‌شود. یکی از مهمترین تکنیک‌های خوشه بندی، خوشه بندی بر مبنای K-means می‌باشد. در این خوشه بندی نمونه‌ها به K خوشه تقسیم می‌شوند. اگر چه الگوریتم K-means راحت و در بسیاری از مسائل سریع می‌باشد، اما دو مشکل اساسی دارد، یکی حساسیت به فرض اولیه و دیگری قرارگرفتن در می‌نیمم محلی می‌باشد. اگر چه با استفاده از الگوریتم‌های K-harmonic means و Fuzzy K-means می‌توان مشکل حساسیت به فرض اولیه را برطرف کرد، اما همچنان مشکل قرار گرفتن در می‌نیمم محلی وجود دارد. دراین تحقیق ما روشی را برای حل مشکل قرار گرفتن در می‌نیمم محلی بر مبنای تکنین شبیه سازی حرارتی ارائه داده‌ایم. K-means (KM) و K-harmonic means (KHM) و Fuzzy K-means (FKM) سه الگوریتمی می‌باشند که در این تحقیق برای تاثیر اثر الگوریتم شبیه سازی حرارتی بر روی خوشه بندی مورد استفاده قرار گرفته‌اند. این الگوریتم‌ها همگی از نوع الگوریتم‌های خوشه بندی بر مبنای مرکز می‌باشند.

شبیه سازی حرارتی همانطور که از نامش پیداست، روشی بر مبنای گرم شدن و دوباره سرد شدن فلزات و رسیدن به می‌نیمم انرژی ساختار کریستالی می‌باشد و به همین خاطر در بیشتر سیستم‌ها برای یافتن می‌نیمم مطلق استفاده می‌شود. این الگوریتم اولین بار بوسیله متروپولیس برای یافتن آرایش تعادلی مجموعه‌ای از اتمها در یک دمای خاص پیشنهاد شد. این الگوریتم از نوع الگوریتم‌های فرا ابتکاری می‌باشد. مزیت اصلی الگوریتم شبیه سازی حرارتی نسبت به بقیه روش‌های فرا ابتکاری همچون الگوریتم‌های جستجوی ممنوع، ژنتیک، شبکه‌های عصبی، بهینه‌سازی مورچه‌ای، توانایی آن در رهایی از می‌نیمم محلی می‌باشد. مساله‌ای که در حل مشکل الگوریتم‌های خوشه بندی می‌تواند موثر باشد.

ما الگوریتم پیشنهادی را بر روی چندین دسته از داده‌های متعارف و مشهور مانند Iris و Wine و Breast cancer اجرا کرده‌ایم و مشاهده کرده‌ایم که با استفاده از الگوریتم شبیه سازی حرارتی توانسته‌ایم مشکلات موجود در الگوریتم‌های خوشه بندی همچون مساله حساسیت به فرض اولیه و قرار گرفتن در می‌نیمم محلی را به خوبی برطرف کنیم و بتوانیم به می‌نیمم مطلق در مقدار تابع هدف دست یابیم.

فهرست مطالب

عنوان **صفحه**

فهرست جداول

عنوان **صفحه**

فهرست شکلها

عنوان **صفحه**

فصل اول

کلیات طرح

ما در جهانی پر از داده زندگی می‌کنیم. هرروزه انسان‌ها با حجم وسیعی از اطلاعات روبه‌رو هستند که باید آنها را ذخیره سازی یا نمایش دهند. یکی از روش‌های حیاتی کنترل و مدیریت این داده‌ها، کلاس بندی[۱] و خوشه بندی[۲] داده‌های با خواص مشابه، درون مجموعه‌ای از دسته‌ها یا خوشه‌ها می‌باشد. امروزه، خوشه بندی نقش حیاتی در روش‌های بازیابی اطلاعات دارد. اساسا سیستم های خوشه بندی همراه با نظارت یا بدون نظارت هستند.

برخلاف کلاس بندی در خوشه بندی، گروه‌ها از قبل مشخص نمی‌باشند و همچنین معلوم نیست که بر حسب کدام خصوصیات گروه بندی صورت می‌گیرد. درنتیجه پس از انجام خوشه بندی باید یک فرد خبره خوشه‌های ایجاد شده را تفسیر کند و در بعضی مواقع لازم است که پس از بررسی خوشه‌ها بعضی از پارامترهایی که در خوشه بندی در نظر گرفته شده‌اند ولی بی‌ربط بوده یا اهمیت چندانی ندارند حذف شده و جریان خوشه بندی از اول صورت گیرد.

هدف نهایی خوشه بندی این است که داده‌های موجود را به چند گروه تقسیم کنند و در این تقسیم بندی، داده‌های گروه‌های مختلف باید حداکثر تفاوت ممکن را به هم داشته باشند و داده‌های موجود در یک گروه باید بسیار به هم شبیه باشند. در خوشه بندی از الگوریتم‌های متعددی استفاده می‌شود که هدف همه این الگوریتم‌ها رسیدن به جواب بهینه می‌باشد.[۱]

---

[۱] Classification
[۲] Clustering

سیستم‌های پیچیده اجتماعی، تعداد زیادی از مسائل دارای طبیعت ترکیباتی[۱] را پیش روی قرار می‌دهند. مسیر کامیون‌های حمل و نقل باید تعیین شود، انبارها یا نقاط فروش محصولات باید جایابی شوند، شبکه‌های ارتباطی باید طراحی شوند، کانتینرها باید بارگیری شوند، رابط‌های رادیویی می‌بایست دارای فرکانس مناسب باشند، مواد اولیه چوب، فلز، شیشه و چرم باید به اندازه‌های لازم بریده شوند، از این دست مسائل بی‌شمارند. تئوری پیچیدگی به ما می‌گوید که مسائل ترکیباتی اغلب پلی‌نومیال[۲] نیستند. این مسائل در اندازه‌های کاربردی و عملی خود به قدری بزرگ هستند که نمی‌توان جواب بهینه آنها را در مدت زمان قابل پذیرش به دست آورد. با این وجود، این مسائل باید حل شوند و بنابراین چاره‌ای نیست که به جواب‌های زیر بهینه[۳] بسنده نمود به گونه‌ای که دارای کیفیت قابل پذیرش بوده و در مدت زمان قابل پذیرش به دست آیند.

چندین رویکرد برای طراحی جواب‌هایی با کیفیت قابل پذیرش تحت محدودیت زمانی قابل پذیرش پیشنهاد شده است. الگوریتم‌هایی هستند که می‌توانند یافتن جواب‌های خوب در فاصله مشخصی از جواب بهینه را تضمین کنند که به آن‌ها الگوریتم‌های تقریبی می‌گویند. الگوریتم‌های دیگری نیز هستند که تضمین می‌دهند با احتمال بالا جواب نزدیک بهینه تولید کنند که به آن‌ها الگوریتم‌های احتمالی گفته می‌شود. جدای از این دو دسته، می‌توان الگوریتم‌هایی را پذیرفت که هیچ تضمینی در ارائه جواب ندارند اما براساس شواهد و سوابق نتایج آن‌ها، به طور متوسط بهترین تقابل کیفیت و زمان حل برای مسئله مورد بررسی را به همراه داشته‌اند. به این الگوریتم‌ها، الگوریتم‌های ابتکاری گفته می‌شود[۲]

الگوریتم‌های ابتکاری عبارتند از معیارها، روش‌ها یا اصولی برای تصمیم‌گیری بین چند گزینه خط‌مشی و انتخاب اثربخش‌ترین برای دستیابی به اهداف مورد نظر. الگوریتم‌های ابتکاری نتیجه برقراری اعتدال بین دو نیاز هستند: نیاز به ساخت معیارهای ساده و در همان زمان توانایی تمایز

---

[۱] combinatorial
[۲] polynomial
[۳] Sub optimal

درست بین انتخاب‌های خوب و بد. برای بهبود این الگوریتم‌ها از اواسط دهه هفتاد، موج تازه‌ای از رویکردها آغاز گردید. این رویکردها شامل الگوریتم‌هایی است که صریحاً یا به صورت ضمنی تقابل بین ایجاد تنوع جستجو (وقتی علائمی وجود دارد که جستجو به سمت مناطق بد فضای جستجو می‌رود) و تشدید جستجو (با این هدف که بهترین جواب در منطقه مورد بررسی را پیدا کند) را مدیریت می‌کنند. این الگوریتم‌ها فرا ابتکاری نامیده می‌شوند. از بین این الگوریتم‌ها می‌توان به موارد زیر اشاره کرد:[۳]

- شبیه‌سازی حرارتی[۱]
- جستجوی ممنوع[۲]
- الگوریتم‌های ژنتیک[۳]
- شبکه‌های عصبی مصنوعی[۴]
- بهینه‌سازی مورچه‌ای یا الگوریتم‌های مورچه[۵]

در این تحقیق ما از الگوریتم شبیه سازی حرارتی جهت مسائل خوشه بندی استفاده خواهیم کرد.[۴]

---

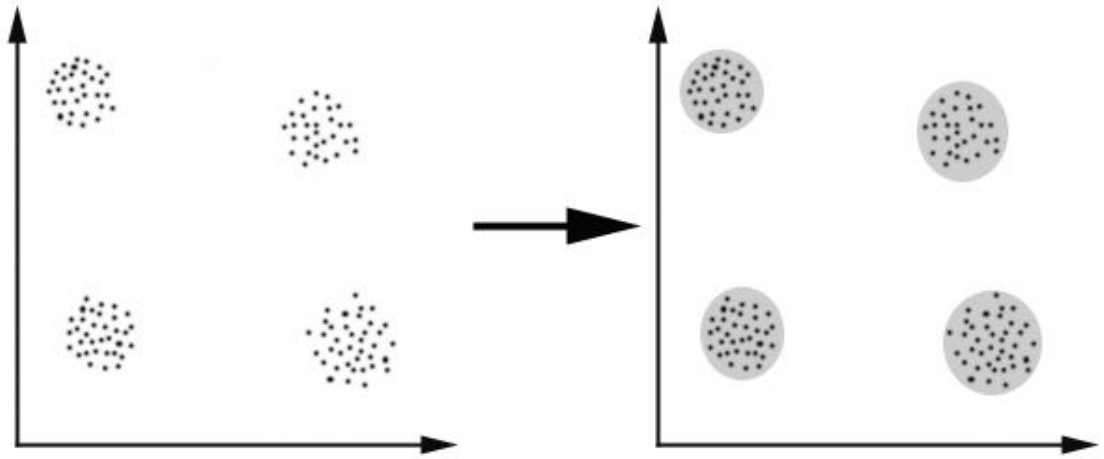
[۱] Simulated annealing (sa)
[۲] Tabu search (ts)
[۳] Genetic algorithms (ga)
[٤] Neural networks
[٥] Ant colony optimization (aco)
[6] Clustering

فصل دوم

خوشه بندی

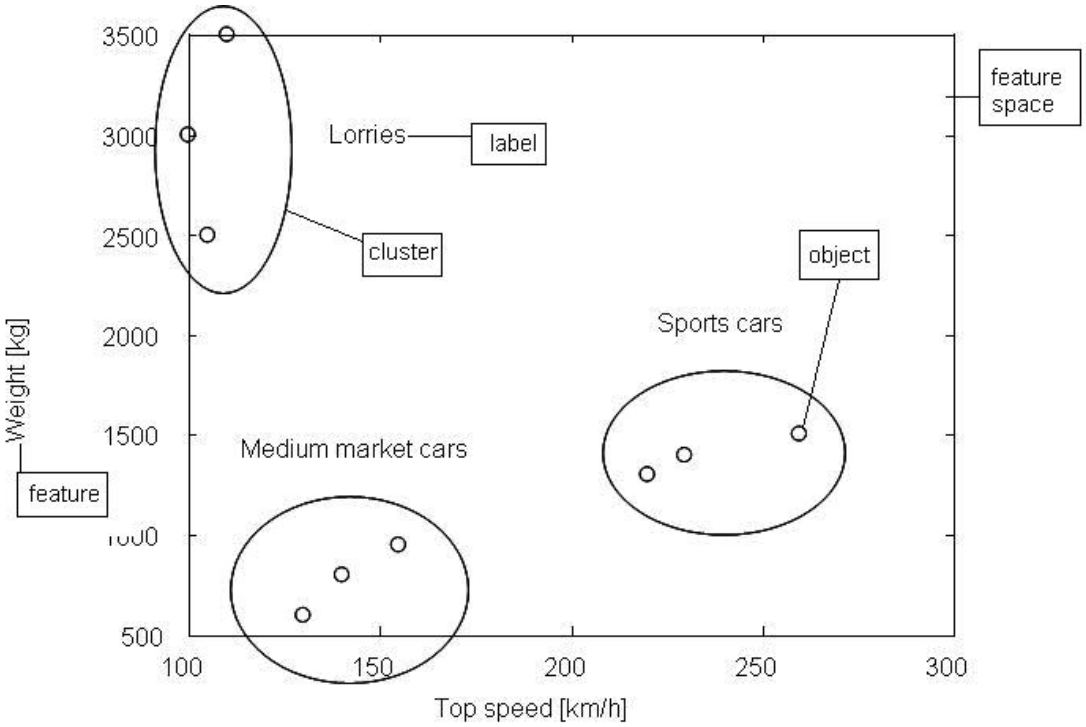
۲-۱ خوشه بندی چیست؟

خوشه بندی یکی از شاخه‌های یادگیری بدون نظارت می‌باشد و فرآیند خودکاری است که در طی آن، نمونه‌ها به دسته‌هایی که اعضای آن مشابه یکدیگر می‌باشند تقسیم می‌شوند که به این دسته‌ها خوشه گفته می شود. بنابراین خوشه مجموعه‌ای از اشیاء می باشد که در آن اعضای مجموعه با یکدیگر مشابه بوده و با اشیاء موجود در خوشه‌های (مجموعه های) دیگر غیر مشابه می‌باشند. برای مشابه بودن می‌توان معیارهای مختلفی را در نظر گرفت. مثلا می‌توان معیار فاصله را برای خوشه بندی مورد استفاده قرار داد و اشیائی را که به یکدیگر نزدیکتر هستند را به عنوان یک خوشه در نظر گرفت که به این نوع خوشه بندی، خوشه بندی مبتنی بر فاصله نیز گفته می‌شود. به عنوان مثالی از خوشه بندی، در شکل ۱ نمونه‌های ورودی در سمت چپ به چهار خوشه مشابه شکل سمت راست تقسیم می‌شوند. در این مثال هر یک از نمونه‌های ورودی به یکی از خوشه‌ها تعلق دارد و نمونه ای وجود ندارد که متعلق به بیش از یک خوشه باشد.



شکل ۱: خوشه بندی نمونه های ورودی

به عنوان یک مثال دیگر شکل ۲ را در نظر بگیرید. در این شکل هر یک از دایره‌های کوچک یک وسیله نقلیه (شیء) را نشان می‌دهد که با ویژگی‌های وزن و حداکثر سرعت مشخص شده‌اند. هر یک از بیضی‌ها یک خوشه می‌باشد و عبارت کنار هر بیضی برچسب آن خوشه را نشان می‌دهد. کل دستگاه مختصات که نمونه‌ها در آن نشان داده شده‌اند را فضای ویژگی می‌گویند.



شکل ۲: خوشه بندی وسایل نقلیه[۵]

همانطور که در شکل می‌بینید وسایل نقلیه به سه خوشه تقسیم شده‌اند. برای هر یک از این خوشه‌ها می توان یک نماینده در نظر گرفت، مثلا می توان میانگین وسایل نقلیه باری را محاسبه کرد و به عنوان نماینده خوشه وسایل نقلیه باری معرفی نمود. در واقع الگوریتم‌های خوشه بندی اغلب بدین گونه‌اند که یک سری نماینده اولیه برای نمونه‌های ورودی در نظر گرفته می‌شود و سپس از روی میزان تشابه نمونه‌ها با این نماینده‌ها، مشخص می‌شود که نمونه به کدام خوشه تعلق دارد و بعد از این مرحله نماینده‌های جدید از روی نمونه‌های متعلق به خوشه محاسبه می شوند و دوباره نمونه‌ها با این نماینده‌ها مقایسه می‌شوند تا مشخص شود که به کدام خوشه تعلق دارند و این کار آنقدر تکرار می شود تا زمانی‌که نماینده‌های خوشه‌ها تغییری نکنند.[۵]

۲-۲ هدف خوشه بندی

هدف خوشه بندی یافتن خوشه‌های مشابه از اشیاء در بین نمونه‌های ورودی می‌باشد، اما چگونه می‌توان گفت که یک خوشه بندی مناسب است و خوشه بندی دیگر مناسب نیست؟ در واقع هیچ معیار مطلقی برای بهترین خوشه بندی وجود ندارد بلکه این بستگی به مساله و نظر کاربر دارد که باید تصمیم بگیرد که آیا نمونه‌ها بدرستی خوشه بندی شده‌اند یا خیر. با این حال معیارهای مختلفی برای خوب بودن یک خوشه بندی ارائه شده است که می توانند کاربر را برای رسیدن به یک خوشه بندی مناسب راهنمایی کند که در قسمت معیارهای کارایی چند نمونه از این معیارها آورده شده است. یکی از مسایل مهم در خوشه بندی انتخاب تعداد خوشه‌ها می باشد. در بعضی از الگوریتم‌ها تعداد خوشه‌ها از قبل مشخص شده است و در بعضی دیگر خود الگوریتم تصمیم می گیرد که داده‌ها به چند خوشه تقسیم شوند.

۲-۳ خوشه‌بندی در مقابل طبقه‌بندی

خوشه بندی با طبقه بندی متفاوت است. در طبقه بندی نمونه‌های ورودی برچسب گذاری شده‌اند ولی در خوشه بندی نمونه‌های ورودی دارای بر چسب اولیه نمی‌باشند و در واقع با استفاده از روشهای خوشه بندی است که داده‌های مشابه مشخص و بطور ضمنی برچسب گذاری می‌شوند. در بسیاری از موارد قبل از عملیات طبقه بندی داده‌ها، یک خوشه بندی روی نمونه‌ها انجام گرفته و سپس مراکز خوشه‌های حاصل را محاسبه می‌کنند و یک بر چسب به خوشه‌ها نسبت می‌دهند و سپس عملیات طبقه بندی را برای نمونه‌های ورودی جدید انجام می‌دهند.در طبقه‌بندی هر داده به یک طبقه (کلاس) از پیشین مشخص شده تخصیص می‌یابد ولی در خوشه‌بندی هیچ اطلاعی از کلاس‌های موجود درون داده‌ها وجود ندارد و به عبارتی خود خوشه‌ها نیز از داده‌ها استخراج می‌شوند. در شکل زیر تفاوت بین خوشه‌بندی و طبقه‌بندی بهتر نشان داده شده است. [۶]