

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



دانشگاه شهید بهشتی

دانشکده علوم ریاضی

گروه آمار

پایان‌نامه‌ی کارشناسی ارشد آمار ریاضی

عنوان

خوشه‌بندی مدل - پایه‌ای به روش پیزی

نگارش

شیما شهبازی

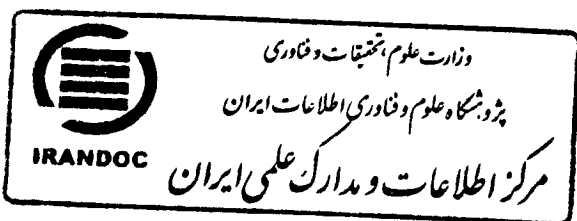
استاد راهنما

دکتر مجتبی خزایی

استاد مشاور

دکتر مسعود البرز

شهریور ۸۹



۱۴۹۵۴۳

۱۳۸۹/۱۰/۲۰

تاریخ
 شماره
 پست
 ۲۹۹۰۱۰

صور تجلسه دفاع از پایان نامه دانشجویان دوره کارشناسی ارشد

ن ۱۹۸۳۹۶۳۱۱۳ اوین

بازگشت به مجوز دفاع شماره ۸۹/۶/۱۹ مورخ ۸۹/۶/۱۹ جلسه هیأت داوران ارزیابی پایان نامه
 خانم شیما شهبازی شماره شناسنامه ۱۷۴ صادره از تهران متولد ۱۳۶۳ دانشجوی آمار ریاضی دوره
 کارشناسی ارشد آمار ریاضی
 با عنوان:

خوشه‌بندی مدل - پایه‌ای به روش بیزی

به راهنمایی:

دکتر مجتبی خزایی

طبق دعوت قبلی در تاریخ ۸۹/۶/۱۹ تشکیل گردید و براساس رأی هیأت داوری و با عنایت به ماده آئین
 نامه کارشناسی ارشد مورخ ۷۵/۱۰/۲۵ پایان نامه مزبور با نمره ۱۹,۲۵ و درجه عالی مورد
 تصویب قرار گرفت.

نام استاد

مرتبه علمی

نام دانشگاه

(۱) استاد راهنما: دکتر مجتبی خزایی

استادیار

شهید بهشتی

(۲) استاد مشاور: دکتر مسعود البرز

استادیار

شهید بهشتی

(۳) داور: دکتر محمدرضا فقیهی

استادیار

شهید بهشتی

(۴) داور: دکتر محمدرضا فریدروحانی

استادیار

شهید بهشتی

نماینده تحصیلات تکمیلی:

کلیه حقوق اعم از چاپ و تکثیر، نسخه برداری، ترجمه، اقتباس و ... از این پایان نامه برای دانشگاه شهید بهشتی محفوظ است. نقل مطالب با ذکر مأخذ بلامانع است.

تقدیم به

پدر و مادر عزیزم

و همه کسانی که دوستشان دارم.

قدردانی

پس از نام و یاد خدا و تشکر از او به خاطر همه‌ی مهربانی‌هایش، بر خود لازم می‌دانم مراتب قدردانی خود را به استاد گرانقدر جناب آقای دکتر مجتبی خزایی، که مرا در تمامی دوره‌ی تحصیلم در دانشگاه شهید بهشتی یاری کردند، اعلام دارم. از آقای دکتر البرز نیز، به خاطر راهنمایی‌های ارزنده‌شان تشکر می‌کنم. همچنین، از اساتید گرانقدر آقای دکتر فقیهی و آقای دکتر فریدروحانی، به سبب حضور در جمع داوران سپاسگزارم.

در پایان از همه‌ی کسانی که مرا در این راه یاری کردند، قدردانی می‌کنم.

شیما شهبازی

شهریور ۱۳۸۹

✓ لوح فشرده‌ی حاوی برنامه‌ها در انتهای پایان‌نامه موجود است.

چکیده

روش‌های خوشه‌بندی سلسله‌مراتبی، از انواع روش‌هایی هستند که جهت خوشه‌بندی اشیاء، براساس میزان عدم تشابه (یا تشابه) بین آن‌ها، به کار می‌روند. در این پایان‌نامه، پس از مرور مختصری بر این روش‌ها، شیوه‌های جدیدتر خوشه‌بندی اشیاء، براساس میزان عدم تشابه بین آن‌ها، معرفی می‌شوند. یکی از مزیت‌های این شیوه‌ها نسبت به روش‌های سلسله‌مراتبی، قابلیت آن‌ها در برآورد عدم حتمیت نتایج خوشه‌بندی است. یکی از این شیوه‌ها، خوشه‌بندی مدل-پایه‌ای اشیاء پس از برآورد پیکربندی آن‌ها در فضای اقلیدسی، به کمک روش‌های مقیاس‌بندی چندبعدی کلاسیک یا بیزی است. روش دیگر، روشی بیزی است که در آن، یافتن پیکربندی اشیاء و خوشه‌بندی آن‌ها به طور هم‌زمان انجام می‌گیرد.

به طور کلی، در روش‌های معرفی شده در این پایان‌نامه، فرض می‌شود که جامعه دارای چگالی نرمال آمیخته است. به این ترتیب، خوشه‌ای کردن بر پایه‌ی یک مدل احتمالی انجام می‌گیرد. همچنین، فرض می‌شود که اندازه‌های مشاهده‌شده از میزان عدم تشابه بین اشیاء، دارای چگالی نرمال بریده‌شده‌اند و با یک رهیافت بیزی، پیکربندی اشیاء برآورد می‌شود. برای انجام این برآورد از روش‌های مونت کارلوی زنجیره‌ی مارکوفی استفاده می‌شود.

این روش‌ها، با استفاده از مطالعات شبیه‌سازی شده مقایسه گردیده‌اند. نتایج حاصل از شبیه‌سازی، گویای یک پیکربندی خوب و خوشه‌بندی معقول با اندازه‌های قابل قبولی از عدم حتمیت در روش‌های جدید است. در پایان، با استفاده از این روش‌ها، یک مجموعه داده‌ی واقعی، تحلیل و نتایج آن ارائه شده است.

واژه‌های کلیدی : خوشه‌بندی سلسله‌مراتبی، عدم تشابه، مقیاس‌بندی چندبعدی، مدل‌های آمیخته، روش‌های مدل-پایه‌ای، روش‌های بیزی، مونت-کارلوی زنجیره‌ی مارکوفی.

پیش‌گفتار

رشد روزافزون حجم اطلاعات و داده‌ها در دنیای امروز، استخراج تفسیر و تحلیل آن‌ها را، با مشکل مواجه کرده است. بر این اساس، روش‌های چندمتغیره‌ی آماری که به کشف اطلاعات مفید از بانک‌های اطلاعاتی منجر می‌شوند، مورد توجه بسیاری از تحلیل‌گران قرار گرفته‌اند. از جمله‌ی این روش‌ها، روش‌های خوشه‌بندی و مقیاس‌بندی چندبعدی هستند که استفاده‌ی روزافزون از آن‌ها در علوم بیولوژیک، فیزیک و علوم اجتماعی، مؤید این مطلب است.

روش‌های خوشه‌بندی، روش‌هایی برای شناسایی گروه‌های مشابه از اشیاء در بین داده‌ها هستند که با استفاده از آن‌ها، حجم وسیعی از داده‌ها، به خوبی سازمان‌دهی و خلاصه می‌شوند. به عنوان مثال، فرض کنید یک مجموعه‌ی بزرگ از صفات اندازه‌گیری شده‌ی تعدادی از بیماران افسرده، در دسترس باشد. تشخیص گروه‌های مجزا در این مجموعه، در صورت وجود، می‌تواند بیانگر انواع مختلف بیماری باشد و در انتخاب یک روش مناسب درمانی، مورد استفاده قرار بگیرد. به این ترتیب خوشه‌بندی می‌تواند نقش مهمی را در درمان ایفا کند. یکی دیگر از مثال‌های کاربرد خوشه‌بندی، خوشه‌بندی مشتریان بر اساس داده‌های برگرفته از تعامل آن‌ها با بنگاه‌های تجاری است. به وسیله‌ی خوشه‌بندی مشتریان می‌توان الگوهای رفتاری آن‌ها را کشف کرد و با تحلیل این الگوها، استراتژی‌های بنگاه را برای هر یک از خوشه‌های مشتریان تدوین کرد. از دیگر کاربردهای این روش، گروه‌بندی افراد جامعه بر اساس صفات مشترک آن‌ها است. بدین منظور رفتارهای متقابل افراد در جامعه، مورد بررسی قرار می‌گیرد و سپس، بر اساس داده‌ها و اطلاعات جمع‌آوری شده، گروه‌های مجزا تعیین می‌شوند. به این ترتیب با تقسیم جامعه به گروه‌های کوچک‌تر، مطالعه‌ی روابط اجتماعی افراد توسط جامعه‌شناسان آسان‌تر می‌شود.

در تحلیل‌های خوشه‌ای، ممکن است که با دو نوع داده سروکار پیدا کنیم. در نوع اول، داده‌ها

شامل مشاهدات یا مقادیر اندازه‌گیری شده‌ی برخی از صفات اشیاء هستند که به صورت n بردار در فضای \mathbb{R}^n ثبت می‌شوند. در نوع دوم، داده‌ها شامل مشاهدات یا مقادیر اندازه‌گیری شده از میزان عدم تشابه (یا تشابه) بین اشیاءند که به صورت یک ماتریس $n \times n$ ثبت می‌شوند. به عنوان مثال، مقادیر اندازه‌گیری شده‌ی عدم تشابه (یا تشابه) روانی بین افراد، در علوم روانشناسی از این نوع‌اند.

متناسب با انواع فوق از داده‌ها، تکنیک‌های مختلف خوشه‌بندی نیز گسترش یافته‌اند. از جمله‌ی این تکنیک‌ها، روش‌های خوشه‌بندی سنتی، مانند انواع روش‌های سلسله‌مراتبی و نیز روش k -میانگین و انواع روش‌های افزایشی، هستند که در فصل اول این پایان‌نامه، مروری سریع بر آن‌ها خواهد شد. از انواع دیگر این تکنیک‌ها، روش‌های پیشرفته‌تری هستند که طی سال‌های اخیر گسترش یافته‌اند. مانند روش خوشه‌بندی مدل-پایه‌ای که توسط اسکات و سیمنز (۱۹۷۱) معرفی و توسط بنفیلد و رفتری (۱۹۹۳)، فرلی و رفتری (۱۹۹۸) و فرلی و رفتری (۲۰۰۲) تکامل یافت. در فصل دوم این پایان‌نامه، در مورد این روش و مزایای استفاده از آن، به طور مفصل بحث می‌شود.

استفاده از روش خوشه‌بندی مدل-پایه‌ای، نیاز به بردار مشاهدات دارد و در حالتی که داده‌ها به صورت ماتریس عدم تشابه بین اشیاءند، قابل استفاده نیست. در این وضعیت، یک راهکار، می‌تواند تبدیل ماتریس عدم تشابه‌ها به یک پیکربندی، شامل n نقطه در فضای \mathbb{R}^n و استفاده از روش خوشه‌بندی مدل-پایه‌ای روی پیکربندی حاصل باشد. خوشبختانه، امروزه، روش‌های مناسبی جهت یافتن پیکربندی اشیاء گسترش یافته‌اند که از آن جمله انواع روش‌های مقیاس‌بندی چندبعدی هستند. روش کلاسیک مقیاس‌بندی چندبعدی، در فصل اول معرفی می‌شود. همچنین، یک رهیافت بیزی از این روش‌ها، که توسط اه و رفتری (۲۰۰۱) ارائه شده است، در فصل سوم معرفی می‌شود.

اه و رفتری (۲۰۰۷)، با ترکیب روش خوشه‌بندی مدل-پایه‌ای و رهیافت بیزی مقیاس‌بندی چندبعدی، روشی جدید را معرفی کردند که به منظور خوشه‌بندی اشیاء، با استفاده از فواصل مشاهده‌شده‌ی بین آن‌ها، به کار می‌رود. این روش نیز، در فصل سوم معرفی می‌شود.

در پایان نتایج حاصل از انواع این روش‌ها روی داده‌های شبیه‌سازی شده و واقعی مقایسه می‌شوند. نتایج حاصل، گویای یک پیکربندی خوب و خوشه‌بندی معقول در انواع روش‌های مدل-پایه‌ای است.

فهرست مندرجات

۱	روش‌های مقدماتی تحلیل خوشه‌ای و مقیاس‌بندی چندبعدی	۱
۱ مقدمه	۱.۱
۲ خوشه‌بندی	۲.۱
۳ فواصل متریک	۱.۲.۱
۵ روش‌های خوشه‌بندی	۲.۲.۱
۱۵ تعیین تعداد خوشه‌ها	۳.۲.۱
۱۵ مقیاس‌بندی چندبعدی	۳.۱
۱۷ مقیاس‌بندی چندبعدی کلاسیک	۱.۳.۱
۱۸ تعیین ابعاد اشیاء	۲.۳.۱
۲۳	خوشه‌بندی مدل-پایه‌ای	۲
۲۳ مقدمه	۱.۲

۲۴ مدل‌های آمیخته و کاربرد آن‌ها در خوشه‌بندی	۲.۲
۲۶ الگوریتم EM و برآورد پارامترهای مدل آمیخته	۳.۲
۳۱ خوشه‌بندی سلسله‌مراتبی مدل-پایه‌ای	۴.۲
۳۱ ۱.۴.۲ تجزیه‌ی طیفی ماتریس واریانس-کوواریانس	
۴۳ انتخاب بهترین مدل	۵.۲
۴۶ خوشه‌بندی مدل-پایه‌ای	۶.۲
۵۴	۳ خوشه‌بندی مدل-پایه‌ای عدم تشابه‌ها	
۵۴ مقدمه	۱.۳
۵۵ مقیاس‌بندی چندبعدی بیزی	۲.۳
۵۶ ۱.۲.۳ مدل و توزیع‌های پیشین	
۶۲ ۲.۲.۳ استنباط پسین	
۶۴ ۳.۲.۳ یک ملاک بیزی جهت انتخاب بعد اشیاء	
۶۹ خوشه‌بندی مدل-پایه‌ای بیزی با کمک فواصل	۳.۳
۷۳ ۱.۳.۳ مدل و توزیع‌های پیشین	
۷۷ ۲.۳.۳ استنباط پسین	
۷۹ ۳.۳.۳ انتخاب بعد اشیاء و تعداد خوشه‌ها	

۸۱ جمع‌بندی	۴.۳.۳
۸۲ شبیه‌سازی	۴.۳
۹۷ مثال واقعی	۵.۳
۱۰۳ نتیجه‌گیری و پیشنهادات	۶.۳

۱۰۸ **A** **قضایا**

۱۱۰ **B** **روش‌های مونت کارلوی زنجیره‌ی مارکوفی MCMC**

۱۱۰ مقدمه

۱۱۱ الگوریتم متروپولیس-هستینگس

۱۱۳ نمونه‌گیری گیبز

۱۱۴ اجرای روش‌های مونت کارلو

۱۱۷ **C** **روش تبدیل پروکراستیز**

۱۱۹ **D** **مساله‌ی تغییر برچسب**

۱۲۲ **E** **برنامه‌های کامپیوتری**

۱۳۴ **واژه‌نامه‌ی فارسی به انگلیسی**

۱۳۶ **نام‌نامه**

لیست اشکال

۵

- ۱.۱ ویژه‌مقدارهای ماتریس A به ازای تعداد ابعاد $14, 13, \dots, 2, 1$ (p مجموعه‌ی داده‌های توالی پروتئین‌ها) ۲۱
- ۲.۱ نمودار پراکنش سه‌بعدی مختصات برآوردشده‌ی 150° پروتئین، با استفاده از ماتریس عدم‌تشابه بین ساختار این تعداد پروتئین ۲۱
- ۱.۲ تاثیر تغییر مولفه‌های تجزیه‌ی ماتریس واریانس-کوواریانس به روی مساحت، شکل و جهت بیضی‌های تراز، وقتی $p = 2$ ۳۳
- ۲.۲ انتخاب یک مدل نرمال آمیخته با سه مولفه‌ی دایره‌ای شکل (a) و یک مدل نرمال آمیخته با دو مولفه که یکی دایره‌ای و دیگری بیضی شکل است روی یک مجموعه از داده‌های شبیه‌سازی شده (b) ۴۶
- ۳.۲ نمودار پراکنش جفت متغیرهای ماکسیم مساحت سلول، ماکسیم یکنواختی و میانگین بافت ۴۹

- ۴.۲ ملاک BIC به ازای تعداد خوشه‌های متفاوت در انواع مدل‌های نرمال آمیخته
 ۴۹ جهت خوشه‌بندی مدل-پایه‌ای (مجموعه‌ی داده‌های بیماران سرطانی)
- ۵.۲ خوشه‌بندی مدل-پایه‌ای بیماران سرطانی با استفاده از مدل نرمال آمیخته با
 ۵۰ دو مولفه‌ی قطری
 ۵
- ۶.۲ نتایج خوشه‌بندی مدل-پایه‌ای بیماران سرطانی و مقایسه‌ی آن با گروه‌های از
 قبل تعیین شده (نقاط مشکلی، مشاهداتی را نشان می‌دهند که نتایج خوشه‌بندی آن‌ها با
 ۵۱ گروه‌های از قبل تعیین شده متفاوت است)
- ۷.۲ خوشه‌بندی مدل-پایه‌ای بیماران سرطانی با استفاده از مدل نرمال آمیخته با
 دو مولفه‌ی بیضی شکل با واریانس‌های برابر (پس از تبدیل باکس کاکس روی متغیر
 ۵۲ ماکسیمم مساحت سلول)
- ۸.۲ نتایج خوشه‌بندی مدل-پایه‌ای بیماران سرطانی و مقایسه‌ی آن با گروه‌های از
 قبل تعیین شده (پس از تبدیل باکس کاکس روی متغیر ماکسیمم مساحت سلول)
 ۵۳
- ۱.۳ نمودار پراکنش سه‌بعدی مختصات برآوردشده‌ی ۱۵۰ پروتئین، با استفاده از
 روش مقیاس‌بندی چندبعدی بی‌زی - خوشه‌بندی مدل-پایه‌ای ساختار برآوردشده
 در چهار خوشه (رنگ‌های قرمز، آبی، قهوه‌ای و سبز که معرف ساختارهای فضایی
 پروتئین‌ها هستند).
 ۷۱
- ۲.۲ ملاک BIC به ازای تعداد خوشه‌های متفاوت در انواع مدل‌های نرمال آمیخته
 ۷۲ جهت خوشه‌بندی مدل-پایه‌ای (مجموعه‌ی داده‌های توالی پروتئین‌ها)

۳.۳	نمودار پراکنش داده‌های شبیه‌سازی شده از چگالی نرمال دو متغیره‌ی آمیخته
۸۳	با سه مولفه
۴.۳	مقدار ملاک C به ازای تعداد خوشه‌های متفاوت در انواع روش‌های
۸۴	خوشه‌بندی کلاسیک
۵.۳	نقاط فرین در مجموعه‌ی داده‌های شبیه‌سازی شده از چگالی نرمال دو
۸۵	متغیره‌ی آمیخته با سه مولفه
۶.۳	سه خوشه‌ی حاصل از اجرای روش‌های تک اتصالی، اتصال میانگین و
	گرانیکه روی کلیدی داده‌های شبیه‌سازی شده (ردیف اول) روی مجموعه‌ی داده‌ها
۸۶	پس از کنار گذاشتن نقاط فرین (ردیف دوم)
۷.۳	انتخاب اشتباه خوشه‌ها در هر یک از روش‌های خوشه‌بندی کلاسیک، با در
۸۷	نظر گرفتن گروه‌های واقعی داده‌های شبیه‌سازی شده
۸.۳	مقادیر ملاک STRESS به ازای $p = ۱, ۲, \dots, ۵$ در روش مقیاس‌بندی
۸۸	چندبعدی کلاسیک (CMDS)
۹.۳	نمودار پراکنش داده‌های شبیه‌سازی شده (a) و داده‌های برآورد شده با استفاده
۸۹	از روش CMDS (b)

- ۱۰.۳ ملاک BIC به ازای تعداد خوشه‌های متفاوت در انواع مدل‌های نرمال آمیخته جهت خوشه‌بندی مدل-پایه‌ای اشیاء با استفاده از مختصات برآوردشده‌ی آن‌ها به روش CMDS ۸۹
- ۱۱.۳ خوشه‌بندی مدل-پایه‌ای اشیاء با استفاده از مختصات برآوردشده‌ی آن‌ها به روش CMDS ۹۰
- ۱۲.۳ برآورد عدم حتمیت خوشه‌های به‌دست‌آمده از روش مدل-پایه‌ای (a) خطا در خوشه‌بندی مدل-پایه‌ای مشاهدات برآوردشده (b) ۹۱
- ۱۳.۳ نمودار پراکنش داده‌های شبیه‌سازی‌شده (a) و داده‌های برآوردشده با استفاده از روش BMDS (b) ۹۲
- ۱۴.۳ ملاک BIC به ازای تعداد خوشه‌های متفاوت در انواع مدل‌های نرمال آمیخته جهت خوشه‌بندی مدل-پایه‌ای اشیاء با استفاده از مختصات برآوردشده‌ی آن‌ها به روش BMDS ۹۳
- ۱۵.۳ خوشه‌بندی مدل-پایه‌ای اشیاء با استفاده از مختصات برآوردشده‌ی آن‌ها به روش BMDS ۹۴
- ۱۶.۳ برآورد عدم حتمیت خوشه‌های به‌دست‌آمده از روش مدل-پایه‌ای (a) خطا در خوشه‌بندی مدل-پایه‌ای مشاهدات برآوردشده (b) ۹۵

- ۱۷.۳ نمودار پراکنش داده‌های شبیه‌سازی شده (a) و داده‌های برآورد شده با استفاده از روش اه و رفتری (b) ۹۷
- ۱۸.۳ برآورد پیکربندی اشیاء و خوشه‌بندی مدل-پایه‌ای آن‌ها با استفاده از روش اه و رفتری ۹۸
- ۱۹.۳ نقاط قرمز رنگ: نمودار پراکنش فواصل مشاهده شده و فواصل برآورد شده با استفاده از روش BMDS. نقاط آبی رنگ: نمودار پراکنش فواصل مشاهده شده و فواصل برآورد شده با استفاده از روش CMDS. (مجموعه‌ی داده‌های کارمندان بانک) . ۱۰۱
- ۲۰.۳ نمودار پراکنش جفت متغیرهای برآورد شده با استفاده از روش BMDS (مجموعه‌ی داده‌های کارمندان بانک) ۱۰۲
- ۲۱.۳ ملاک BIC به ازای تعداد خوشه‌های متفاوت در انواع مدل‌های نرمال آمیخته جهت خوشه‌بندی مدل-پایه‌ای (مجموعه‌ی داده‌های کارمندان بانک) ۱۰۴
- ۲۲.۳ خوشه‌بندی مدل-پایه‌ای کارمندان بانک با استفاده از مختصات برآورد شده‌ی آن‌ها به روش BMDS ۱۰۴

لیست جداول

۱۱	مقادیر پارامترها در روش بتای انعطاف پذیر	۱.۱
۳۴	انواع مدل های احتمالی نرمال آمیخته	۱.۲
۸۷	ملاک $P_k^{(1)}$ و ملاک $P_k^{(2)}$ به ازای $k = 1, 2, 3$	۱.۳
۹۱	...	عدم حتمیت خوشه بندی، پس از برآورد مختصات اشیاء به روش CMDS	۲.۳
۹۴	...	عدم حتمیت خوشه بندی، پس از برآورد مختصات اشیاء به روش BMDS	۳.۳
۹۶	$p = 1, 2, 3, 4$ به ازای MDSIC و LR_p , CMDS, BMDS STRESS	۴.۳
۱۰۰	ملاک LR_p و ملاک MDSIC به ازای $p = 1, 2, \dots, 9$	۵.۳