بسم الله الرحمن الرحيم

٤٠٨٥٨

دانشگاه تربیت مدرس

دانشکده علوم پایه

پایان نامه دوره کارشناسی ارشد آمار

کاربرد گرافهای تصادفی بازه‌ای در تحلیل خوشه‌ای

توسط

فرانک گودرزی

استاد راهنما

دکتر محمدقاسم وحیدی اصل

استاد مشاور

دکتر محسن محمدزاده

اسفند ماه ۱۳۸۰

تأییدیه اعضای هیأت داوران حاضر در جلسه دفاع از پایان نامه کارشناسی ارشد

اعضای هیئت داوران نسخهٔ نهایی پایان نامه خانم/ آقای فرانک گودرزی

تحت عنوان: **کاربرد گرافهای تصادفی بازه‌ای در تحلیل خوشه‌ای**

را از نظر فرم و محتوا بررسی نموده و آنرا برای اخذ درجه کارشناسی‌ارشد مورد تایید قرار دادند.

اعضای هیأت داوران	نام و نام خانوادگی	رتبهٔ علمی	امضاء
۱- استاد راهنما	آقای دکتر محمدقاسم وحیدی‌اصل	دانشیار	
۲- استاد مشاور	آقای دکتر محسن محمدزاده	استادیار	
۳- استاد ناظر	آقای دکتر عین‌اله پاشا	دانشیار	
۴- استاد ناظر	آقای دکتر عباس گرامی	استادیار	
۵- نماینده تحصیلات تکمیلی	آقای دکتر عین‌اله پاشا	دانشیار	

بسمه‌تعالی

آیین‌نامهٔ چاپ پایان‌نامه (رساله)های دانشجویان دانشگاه تربیت مدرّس

نظر به اینکه چاپ و انتشار پایان‌نامه (رساله)های تحصیلی دانشجویان دانشگاه تربیت مدرّس، مبیّن بخشی از فعالیتهای علمی - پژوهشی دانشگاه است بنابراین به منظور آگاهی و رعایت حقوق دانشگاه، دانش‌آموختگان این دانشگاه نسبت به رعایت موارد ذیل متعهد می‌شوند:

مادّهٔ ۱ در صورت اقدام به چاپ پایان‌نامه (رساله)ی خود، مراتب را قبلاً به طور کتبی به «دفتر نشر آثار علمی» دانشگاه اطلاّع دهد.

مادّهٔ ۲ در صفحهٔ سوم کتاب (پس از برگ شناسنامه)، عبارت ذیل را چاپ کند:

«کتاب حاضر، حاصل پایان‌نامهٔ کارشناسی ارشد/رسالهٔ دکتری نگارنده در رشتهٔ آمار *است که در سال* ۱۳۸۰ *در دانشکدهٔ* علوم پایه *دانشگاه تربیت مدرّس به راهنمایی* ~~*سرکار خانم*~~*/جناب آقای دکتر* محمدقاسم وحیدی *، مشاورهٔ* ~~*سرکار خانم*~~*/جناب آقای دکتر* محسن محمدزاده *و مشاورهٔ سرکار خانم/جناب آقای دکتر* — *از آن دفاع شده است.»*

مادّهٔ ۳ به منظور جبران بخشی از هزینه‌های انتشارات دانشگاه، تعداد یک درصد شمارگان کتاب (در هر نوبت چاپ) را به «دفتر نشر آثار علمی» دانشگاه اهدا کند. دانشگاه می‌تواند مازاد نیاز خود را به نفع مرکز نشر در معرض فروش قرار دهد.

مادّهٔ ۴ در صورت عدم رعایت مادّهٔ ۳، ۵۰٪ بهای شمارگان چاپ‌شده را به عنوان خسارت به دانشگاه تربیت مدرّس، تأدیه کند.

مادّهٔ ۵ دانشجو تعهد و قبول می‌کند در صورت خودداری از پرداخت بهای خسارت، دانشگاه می‌تواند خسارت مذکور را از طریق مراجع قضایی مطالبه و وصول کند؛ به علاوه به دانشگاه حق می‌دهد به منظور استیفای حقوق خود، از طریق دادگاه، معادل وجه مذکور در مادّهٔ ۴ را از محلّ توقیف کتابهای عرضه‌شدهٔ نگارنده برای فروش، تأمین نماید.

مادّهٔ ۶ اینجانب فرانک گودرزی دانشجوی رشته آمار مقطع کارشناسی ارشد تعهد فوق و ضمانت اجرایی آن را قبول کرده، به آن ملتزم می‌شوم.

نام و نام‌خانوادگی: فرانک گودرزی

تاریخ و امضا: ۸۱/۱/۲۰

تقدیم به:

مادر عزیزم،

که دستهای صبورش، سپرد مرا
به دست موج محبت، در آستانه‌ی آب
به پاس آنهمه گرمی
که بر برودت قطبی سینه ام پاشید

قدردانی

با تشکر و سپاس از استاد عالیمقام و گرانقدر ، جناب آقای دکتر محمد قاسم وحیدی اصل که قبول زحمت فرموده و راهنمایی این رساله را عهده‌دار شدند و هرگز از راهنماییهای سودمند در جهت رفع نواقص و بهبود و تکمیل این رساله دریغ نفرموده و در تمام مراحل . صمیمانه مرا یاری نمودند.

همچنین از استاد گرانمایه جناب آقای محسن محمّد زاده که مسئولیت مشاوره این رساله را بر عهده داشته‌اند نهایت تشکر و سپاس را دارم .

به جاست از جناب آقای دکتر عین اللّه پاشا و جناب آقای عباس گرامی که قبول زحمت نموده و داروی این رساله را به عهده گرفته‌اند . قدردانی نمایم .

کاربرد گرافهای تصادفی بازه‌ای در تحلیل خوشه‌ای

چکیده

در این پایان نامه، دو مدل برای گرافهای تصادفی بازه‌ای مطالعه می‌شوند. اولین مدل ایستا است: برای هر عدد صحیح مثبت n، تنها یک فضای احتمال برای گرافهای بازه‌ای روی n رأس برچسب گذاری شده در نظر گرفته می‌شود. سپس یک مدل تکاملی برای چنین گرافهای تصادفی بازه‌ای بسط داده می‌شود. نتایج دقیق و حدی برای توزیع متغیرهای تصادفی مربوط به همبندی گراف تصادفی بازه‌ای اخیر مورد بحث قرار می‌گیرند.

اینک در چارچوب تحلیل خوشه‌ای، فرض کنید n داده در فضای k بعدی اقلیدسی در دست داریم و می‌خواهیم آنها را در معرض یک الگوریتم تحلیل خوشه‌ای قرار دهیم. سؤال عمده‌ای که باید پاسخ داده شود این است که آیا خوشه‌های حاصل یک تفسیر "علّی" دارند یا صرفاً پیامدهای یک نوسان "تصادفی" هستند.

در این پایان نامه، خواص مجانبی تعدادی آزمون ترکیبیاتی بالقوه سودمند بر اساس نظریه گرافهای تصادفی بازه‌ای توصیف می‌شوند. به کمک برخی نتایج عددی مقدماتی کاربرد ممکن آنها به عنوان روشی برای پاسخ به پرسش بالا شرح داده می‌شود. سپس مقایسه‌هایی از اثر مجانبی یک رده از این آزمونها ارائه می‌شوند. به عنوان یک توضیح خاص از کاربردهای ممکن، آشکارسازی آمیزه‌های توزیعهای احتمال مورد بحث قرار گرفته و چند مثال عددی ارائه می‌شود.

واژه‌های کلیدی : گرافهای تصادفی بازه‌ای. تحلیل خوشه‌ای، مدلهای آمیخته، آشکارسازی

فهرست مندرجات

فصل ۱

مروری بر تحلیل خوشه‌ای

۱.۱ مقدمه

تحلیل خوشه‌ای برای حل مسئله‌ای طرح شده است که درآن با در دست داشتن نمونه ای از n فرد و اندازه گیری p متغیر بر روی هر فرد، می‌توان افراد را در رده‌هایی گروه بندی نمود که افراد مشابه در داخل یک رده قرار گیرند. به عبارت دیگر کاربرد انواع روشها و فنها برای کشف ساختار درونی مجموعه‌ای از مشاهدات را تحلیل خوشه‌ای می‌نامند. اصطلاح خوشه‌بندی کردن را مترادف با طبقه‌بندی عددی و رده‌بندی می‌گیرند. این روشها کاملاً عددی هستند و تعداد رده‌ها اغلب مشخص نیست. واضح است که این مسئله مشکلتر از مسئله تحلیل ممیزی است، زیرا در تحلیل ممیزی گروهها از اول مشخص‌اند. دلایل زیادی را می‌توان برای نشان دادن ارزشمندی تحلیل خوشه‌ای ارائه داد.

اولاً تحلیل خوشه‌ای می‌تواند در پیدا کردن گروههای واقعی کارساز باشد. به عنوان مثال، در درمان بیماریهای روانی، اختلاف نظر زیادی برای رده‌بندی بیماران افسرده وجود دارد و تحلیل خوشه‌ای در تعریف گروههای واقعی مورد استفاده قرار می‌گیرد.

دوم اینکه تحلیل خوشه‌ای می‌تواند برای کاهش داده‌ها مفید باشد. به عنوان مثال تعداد زیادی از شهرها به طور بالقوه می‌توانند به صورت بازار آزمایشی برای یک محصول جدید به کار روند، ولی با توجه به محدودیت امکانات فقط آزمایش در

تعداد کمی از شهرها امکان پذیر است. حال اگر شهرها را بتوان به تعداد کمتری از گروههای مشابه گروهبندی کرد، آن وقت یک شهر از هر گروه می تواند به عنوان بازار آزمایشی به کار رود.

از طرف دیگر تحلیل خوشه‌ای ممکن است گروههای غیر قابل انتظاری را ایجاد کند. در این صورت نتیجهٔ حاصل بیانگر روابط جدیدی خواهد بود که باید مورد بررسی قرار گیرند. تحلیل خوشه‌ای در بسیاری از رشته‌ها از جمله باستان شناسی، جغرافیا، گیاه شناسی، زمین شناسی، جانور شناسی، زیست شناسی و انسان شناسی مورد استفاده قرار می‌گیرد.

۲.۱ نمادهای پایه‌ای و تعریفها

فرض کنید مجموعهٔ $I = \{I_۱, I_۲, \ldots, I_n\}$، n فرد از یک جامعهٔ فرضی Π_I را نشان دهد. به طور ضمنی فرض می شود که یک مجموعه از وجوه یا مشخصه‌های $C = (C_۱, C_۲, \ldots, C_p)^T$ وجود دارد به گونه‌ای که قابل مشاهده‌اند و هر فرد در I آن را داراست. اصطلاح قابل مشاهده برای مشخصه‌هایی به کار می‌رود که هم داده کمی و هم داده کیفی را به دست می‌دهد، اگر چه در این پایان نامه بیشتر داده‌های کمی تحت عنوان اندازه‌ها به کار برده می‌شوند. اندازه i امین مشخصهٔ فرد I_j را با نماد $X_{i,j}$ نشان داده و فرض می‌کنیم بردار $p \times ۱$، $X_j = \{X_{i,j}\}$، $i = ۱, ۲, \ldots, p$ و $j = ۱, ۲, \ldots, n$ چنین اندازه‌هایی را نشان دهد. از این رو متناظر هر مجموعه افراد I، یک مجموعه از اندازه‌های بردارهای $p \times ۱$ به صورت $X = \{X_۱, X_۲, \ldots, X_n\}$ وجود دارد که مجموعهٔ I را توصیف می کند. مجموعهٔ X می‌تواند به عنوان n نقطه در فضای اقلیدسی p بعدی، E_p در نظر گرفته شود.

۳.۱ مسئلهٔ خوشه‌بندی

فرض کنید m یک عدد صحیح کمتر از n باشد. بر اساس داده‌های مشتمل در مجموعه X، مسئله خوشه‌بندی عبارت است از تعیین m خوشه (زیر مجموعه) از افراد

در I مانند Π_1، Π_2، $\cdots$، Π_m است به گونه‌ای که I_i به یک و تنها یک زیر مجموعه متعلق باشد و افراد منتسب به یک خوشه متشابه و افراد خوشه‌های مختلف متفاوت (غیر متشابه) می‌باشند.
یک راه حل برای مسئله خوشه‌بندی معمولاً تعیین افرازی است که در یک ملاک بهینگی صدق می‌کند. این ملاک بهینگی ممکن است بر حسب یک رابطه تابعی داده شود که سطوح مرغوبیت افرازهای گوناگون یا گروهبندی‌ها را منعکس می‌کند. این رابطه تابعی اغلب یک تابع عینی نامیده می‌شود. برای مثال، مجموع مربعات درون گروهی ممکن است به عنوان یک تابع عینی به کار رود.
برای "حل" مسئلهٔ خوشه‌بندی، لازم است عبارات "تشابه" و "تفاوت" در یک شیوه کمی تعریف شوند. بنابراین جملهٔ دو فرد I_j و I_k متفاوت هستند به چه معنی است؟
یک پاسخ به مسئله چه بسا این است که شخصی i امین و j امین فرد را به خوشه یکسانی نسبت دهد اگر "فاصله" بین نقاط X_i و X_j "به اندازه کافی کوچک" باشد و به خوشه‌های متفاوت نسبت دهد اگر "فاصله" بین X_i و X_j "به اندازه کافی بزرگ" باشد.

۴.۱ تابعهای فاصله

تعریف ۱.۴.۱: تابع حقیقی مقدار نامنفی $d(.,.)$ یک تابع فاصله (متریک) نامیده می شود هرگاه برای بردارهای X_i، X_j و X_k در فضای اقلیدسی E_p داشته باشیم:

(۱) به ازای هر X_i و X_j در E_p ، $d(X_i, X_j) \geq 0$

(۲) $d(X_i, X_j) = 0$ اگر و تنها اگر $X_i = X_j$

(۳) $d(X_i, X_j) = d(X_j, X_i)$

(۴) $d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)$

برای X_i و X_j معین، مقدار $d(X_i, X_j)$ فاصله بین X_i و X_j نامیده می‌شود و فاصله بین I_i و I_j نسبت به مشخصه‌های انتخاب شده $C = (C_1, C_2, \ldots, C_p)^T$ به طور هم ارز با نماد $d(I_i, I_j)$ نشان داده می‌شود.

مثالهایی از چند تابع فاصله مشهور و مفید در زیر داده می‌شود .

نام	شکل
۱. اقلیدسی	$d_2(X_i, X_j) = \left[\sum_{k=1}^{p}(X_{ki} - X_{kj})^2\right]^{\frac{1}{2}}$
۲. نرم l_1	$d_1(X_i, X_j) = \left[\sum_{k=1}^{p} \mid X_{ki} - X_{kj} \mid\right]$
۳. نرم ـ سوپ	$d_\infty(X_i, X_j) = \sup_{k=1,2,\ldots,p}\{\mid X_{ki} - X_{kj} \mid\}$
۴. نرم l_p	$d_p(X_i, X_j) = [\sum_{k=1}^{p} \mid X_{ki} - X_{kj} \mid^p]^{\frac{1}{p}}$
۵. ماهالانوبیس	$D^2(X_i, X_j) = (X_i - X_j)^T \sum^{-1}(X_i - X_j)$

متریک اقلیدسی، متریک خیلی مشهور و رایجی است. متریک قدر مطلق برای بررسی محاسباتی آسان است. نرم ـ سوپریمم همچنین از نظر محاسباتی ساده است، ولیکن مستلزم یک روش رتبه‌بندی است. نرم l_p شامل توابع فاصله ۱، ۲، ۳ به ترتیب به عنوان حالتهای خاصی برای p برابر با ۲، ۱ و ∞ است.
متریک ماهالانوبیس اغلب به عنوان فاصله اقلیدسی تعمیم یافته تلقی می‌شود، که در آن ماتریس $\sum$ ماتریس پراکنش درونی مشاهدات است. فاصله ماهالانوبیس تحت هر تبدیل خطی نامنفرد ناوردا است.

۵.۱ اندازه‌های تشابه

n اندازهٔ X_1، X_2، $\cdots$، X_n را می‌توان برحسب ماتریس داده‌های $p \times n$ نشان داد.

$$X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pn} \end{pmatrix} = (X_1, X_2, \ldots, X_n)$$

به علاوه فاصله‌های دو به دوی $d(X_i, X_j)$ را می‌توان بر حسب ماتریس فاصله