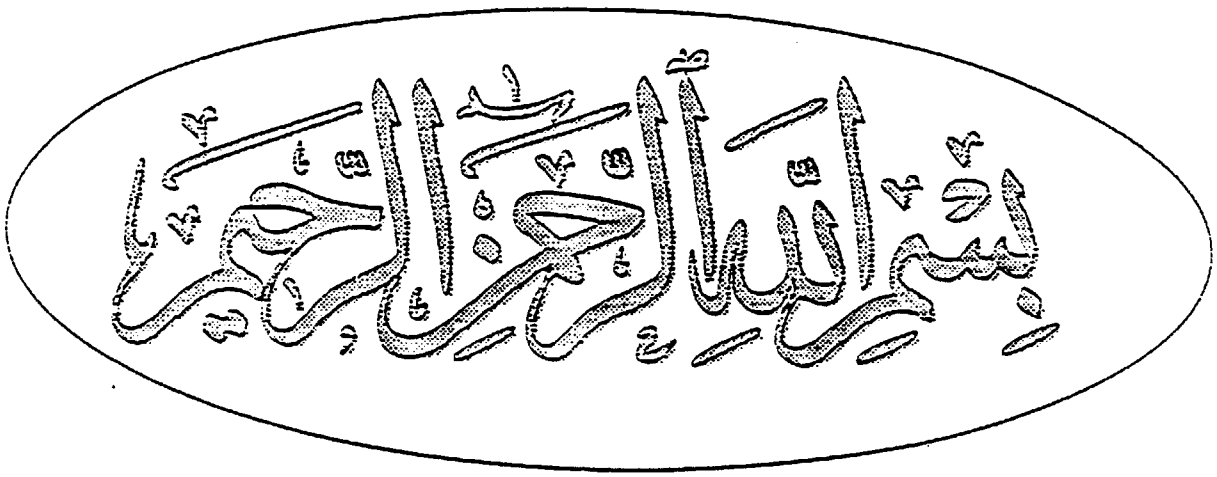

۱۳۷۹ /۱۲/ ۲۰

دانشگاه شهید بهشتی

دانشکده علوم ریاضی

گروه آمار

پایان نامهٔ:

جهت اخذ درجه کارشناسی ارشد آمار

موضوع :

تحلیل خوشه‌ای براساس نظریهٔ گرافهای تصادفی

استادراهنما:

جناب آقای دکتر محمدقاسم وحیدی‌اصل

استاد مشاور:

جناب آقای دکتر علی عمیدی

نگارش:

بابک بابادی

اسفند ۱۳۷۸

بسمه تعالی

تاریخ .....
شماره .....
پیوست .....

دانشکده علوم ریاضی

صورتجلسه دفاع از پایان نامه

جلسه هیأت داوران ارزیابی پایان نامه آقای بابک بابادی

به شماره شناسنامه ۹۳۸ صادره از اهواز متولد ۱۳۵۲ دانشجوی دوره

کارشناسی ارشد ناپیوسته رشته آمار محض

با عنوان : تحلیل خوشه‌ای بر اساس نظریه گرافهای تصادفی

به راهنمائی آقای دکتر محمدقاسم وحیدی اصل طبق دعوت قبلی در تاریخ ۷۸/۱۲/۷

تشکیل گردید و براساس رأی هیأت داوری و با عنایت به ماده ۲۰ آئین نامه کارشناسی

ارشد مورخ ۷۳/۱۰/۲۵ پایان نامه مزبور با نمره ۱۸/۵ هیجده و نیم و درجه عالی

مورد تصویب قرار گرفت .

۱- آقای دکتر محمدقاسم وحیدی اصل	استاد راهنما
۲- آقای دکتر علی عمیدی	استاد مشاور
۳- آقای دکتر خلیل شفیعی	استاد داور
٤- آقای دکتر گنجعلی	استاد داور
۵-	

تقدیم به پدر و مادرم

سپاس و تشکر

رسالهٔ حاضر که طی مدت تقریبی یک سال گردآوری شده، حاصل تلاش اینجانب و کلیه اساتید و دوستانی است که در طول این مدت، بنده را یاری نموده‌اند. در ابتدا لازم است از استاد راهنمای گرامی جناب آقای دکتر وحیدی‌اصل که با ارائه نظرات و پیشنهادات مفید، جهت تکمیل هر چه بهتر مطلب، اینجانب را یاری نموده‌اند، تشکر و قدردانی نمایم. همچنین از اساتید محترم، آقایان دکتر عمیدی به عنوان استاد مشاور، دکتر شفیعی و دکتر گنجعلی به عنوان استاد داور که با مطالعهٔ اولیه رساله نقایص آن را گوشزد نموده‌اند، کمال تشکر را دارم.

از آقایان دکتر مسعود یارمحمدی، محمد طاهری، فتا جعفری‌زاده، خانم مقدم و خصوصاً خانم ناهید رضایی و کلیه دوستان و عزیزانی که درپیشرفت این پایان‌نامه صمیمانه همکاری داشته‌اند، کمال تشکر و قدردانی را دارم. در انتها نیز بر خود واجب می‌دانم از پدر و مادر عزیز که در طی دوران طولانی تحصیل تمام مشکلات را به جان خریده و شرایط لازم برای ادامه تحصیلم را فراهم نمودند، تشکر و قدردانی کرده و این اثر ناچیز را به آنها تقدیم می‌نمایم.

بابک بابادی

فهرست مطالب

چکیده

این رساله درختواره‌نگارهای شاخص‌دار تصادفی تولید شده از الگوریتمهای سلسله مراتبی انباشتی را تحت فرضهای عدم رده‌بندی مربوط به iid عدم مشابهت مورد مطالعه قرار می‌دهد.

آزمونهای جدیدی برای قابل رده‌بندی بودن تهیه شده است. این آزمونها بر اساس متغیرهایی تصادفی از درختواره‌نگارهای شاخص‌دار، مانند دنبالهٔ شاخصها، زمان بقای مجموعه‌های تک عضوی، فاصلهٔ فرامتریک بین دو شیء مفروض یا اندازه‌های خوشه‌ای تعریف می‌شوند. برای یک درختواره شاخص‌دار که به وسیله روش تک پیوندی بر روی iidهای عدم مشابهت تولید شده است. توزیع دقیق و مجانبی این متغیرهای تصادفی محاسبه شده است. برای روشهای پیوندی کامل و پیوند میانگین نیز یک توصیف جزئی از توزیعهای مجانبی بیان این متغیرهای تصادفی و با توزیعهای مجانبی متناظر در روش تک پیوندی مقایسه می‌شوند. اثباتها اساساً به یک تئوری از گرافهای تصادفی مربوط می‌شود.

فرمولهای دقیق برای مقادیر بزرگ قابل استفاده نیست و از نتایج مجانبی به جای آنها استفاده می‌شود. البته سؤالاتی در مورد خصوصیات مجانبها مطرح می‌شود. نتایج ما حاصل از گرافهای تصادفی است و در بیشتر قضیه‌های حدی از تخمین پواسن استفاده می‌شود. این نتایج برای بیان آزمونهایی از عدم رده‌بندی در مقابل فرضهای دیگر که نشان دهندهٔ وجود نوعی از افراز در شیءها است، استفاده می‌شود. در پایان نیز با استفاده از این آزمونهای جدید امکان رده‌بندی آبهای رودخانهٔ کارون در استان خوزستان را مورد بررسی و تجزیه و تحلیل قرار می‌دهیم.

پیشگفتار

روشها و الگوریتمهای مختلفی برای رده‌بندی کردن یک مجموعهٔ متناهی از مشاهدات وجود دارد. ورودیهای الگوریتمها می‌توانند عدم مشابهتها یا مقادیر یک مجموعه از متغیرهای اندازه گیری شده بر روی یک شیء باشند. خروجیها، ساختارهای رده‌بندی شده‌ای چون افزار یا درختواره‌نگارهای شاخص‌دار هستند. حتی اگر داده‌ها به طور آشکار دارای ساختار رده‌بندی شده‌ای نباشد، اغلب الگوریتمها این ساختار را برای داده‌ها به وجود می‌آورند. بنابراین برای تصمیم گیری، این مسئله مهم است که آیا خروجیهای حاصل از یک الگوریتم رده‌بندی، به یک خوشه‌بندی واقعی از داده‌ها اشاره می‌کند؟

ایدهٔ کلی بر تعریف دقیق فرضهای مناسبی از عدم رده‌بندی مطابق با برخی توزیعهای احتمالاتی استوار است. ابتدا باید توزیع احتمال برخی صفات خروجیها را مورد مطالعه قرار دهیم، آنگاه خروجیهای مشاهده شده را می‌توان با یک آستانهٔ مناسب بر پایهٔ یک توزیع، مورد مقایسه قرارداد و فرضهای عدم رده‌بندی را قبول یا رد کرد. رد فرض به معنای این است که خروجیها، به ندرت از نمونه‌های تصادفی (شاخصی برای قابلیت رده‌بندی داده‌ها) می‌باشند. البته ذکر این نکته ضروری است که تعریف یک آزمون بدون در نظر گرفتن شقهای دیگر فرض صفر، ما را به یک آزمون با توان کم رهنمون می‌نماید.

نمایشهای متفاوت از عدم رده‌بندی و انواع متفاوت آزمونهای رده‌بندی را می‌توان در [۳۲و۳۱]، [۶و۵و۴]، [۲۰و۱۹و۱۸] و [۲۴] و تعاریف متفاوت از توزیعهای احتمال مربوط به رده‌بندی تصادفی را در [۱۵]، [۳۰]، [۳۳]، [۲۰و۱۹] و [۶] پیدا کرد.

در این رساله الگوریتمهای سلسله مراتبی انباشتی را برای الگوریتمهای تک پیوندی [۳۷و۱۳]، پیوندی

کامل [۳۹] و پیوند میانگین در نظر گرفته‌ایم[۴۱]. این روشها یک درختواره‌نگار شاخص‌دار را از یک ماتریس که تفاوت دو به دوی اشیا را نشان می‌دهد، تولید می‌کنند. درختواره‌نگارهای شاخص‌دار به طور مستقل به وسیله هارتیگان [۱] [۳۳]، جاردین و سیبسون [۲] [۲۶] و جانسون [۳] [۲۸] تهیه شده‌اند.

در فصل اول این رساله به مفاهیم اساسی از تحلیل خوشه‌ای، انواع روشهای خوشه‌بندی و تعاریفی از فاصلهٔ (مشابهت) بین داده‌ها به عنوان ابزاری جهت رده‌بندی، همچنین بررسی سه روش تک پیوندی، پیوندی کامل و پیوند میانگین جهت خوشه‌بندی یک سری داده اشاره شده است.

در پایان نیز تعدادی از مشکلات تحلیل خوشه‌ای بیان شده است.[۴۲]

در فصل دوم ابتدا تعاریفی از نظریهٔ گرافهای تصادفی بیان می‌شود. سپس مدلهای متفاوت گرافهای تصادفی مانند $G(n,P(edge)=p)$ و $G(n,M)$ تعریف و رابطهٔ بین آنها مورد بررسی قرار می‌گیرد. در پایان نیز خصوصیتهای مهمی مانند وجود درختها، دورها، همبندی، همچنین زمان شکل گیری آنها در مدلهای ذکر شده در بالا در غالب قضیه‌هایی بیان شده است.

در فصل سوم تحلیل احتمالاتی الگوریتمها و بررسی رفتار یک الگوریتم با محاسبهٔ توزیع احتمال خروجیها آن در حالت تصادفی بودن ورودیها صورت می‌پذیرد. در ابتدا تعاریفی از متغیرهای تصادفی مناسب، مربوط به آزمونهای آماری عدم رده‌بندی مشاهدات بیان می‌شود. در بخشهای بعدی (۳ و ۴) توزیعهای دقیق و مجانبی برای تمام متغیرهای تصادفی در حالتی از یک درختواره‌نگار تولید شده به وسیله روش پیوندی بر پایهٔ تعاریفی از نظریهٔ گراف [۷] است. در بخش آخر نیز آزمونهایی برای غیرقابل رده‌بندی بودن در مقابل فرضهای دیگر که نشان دهندهٔ وجود افراز در شیءها می‌باشد، مطرح می‌شود.

---

۱) Hartigan
۲) Jardin and Sibson
۳) Johnson

در فصل چهارم، نحوهٔ رده‌بندی آبهای رودخانهٔ کارون در استان خوزستان با استفاده از دو فاکتور یا هدایت الکتریکی و $EC \times ۱۰^{۶}$ و S.A.R که معرف نسبت جذب سدیم آب می‌باشد، به عنوان کار عملی انجام گرفته است.

در پایان برنامه‌های کامپیوتری مربوط به کار عملی به عنوان ضمیمه و همچنین واژه‌نامه و فهرست مراجع را آورده‌ایم.

مسلماً این رساله دارای نواقص و کاستی‌های فراوانی است که امیدوارم استادان و دانشجویان عزیز با تذکر خود مرا در رفع آنها یاری فرمایند.

فصل اول

مروری بر تحلیل خوشه‌ای