



۴۱۱۳

مرکز اطلاعات و آرکایو علمی ایران  
تسلیت در آرک

# دانشگاه شهید بهشتی

دانشکده علوم ریاضی

گروه آمار

پایان نامه:

جهت اخذ درجه کارشناسی ارشد آمار

موضوع:

تحلیل خوشه‌ای بر اساس نظریه گرافهای تصادفی

استاد راهنما:

جناب آقای دکتر محمد قاسم وحیدی اصل

استاد مشاور:

جناب آقای دکتر علی عمیدی

۸۹۸۶

نگارش:

بابک بابادی

اسفند ۱۳۷۸

۳/۱/۱۳

تاریخ .....  
شماره .....  
پیوست .....

شهرت  
شهرت

دانشکده علوم ریاضی

صور تجلسه 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 راهنمایی آقای دکتر محمد قاسم وحیدی، اصل طبق دعوت قبلی در تاریخ ۷/۱۲/۷۸

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

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

مورد تصویب قرار گرفت .

استاد راهنما  
استاد مشاور  
استاد داور  
استاد داور

۱- آقای دکتر محمد قاسم وحیدی اصل

۲- آقای دکتر علی عمیدی

۳- آقای دکتر خلیل شفیعی

۴- آقای دکتر گنجعلی

# تقدیم به پدر و مادر

## سپاس و تشکر

رساله 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 وحیدی اصل 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 مطالعه اولیه رساله نقایص آن را گوشزد نموده‌اند، کمال تشکر را دارم.

از آقایان دکتر مسعود یارمحمدی، محمد طاهری، فتا جعفری زاده، خانم مقدم و خصوصاً خانم ناهید رضایی و کلیه دوستان و عزیزانی که در پیشرفت این پایان نامه 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 تحصیل را فراهم نمودند، تشکر و قدردانی کرده و این اثر ناچیز را به آنها تقدیم می‌نمایم.

بابک بابادی

# فهرست مطالب

۱	پیشگفتار
۴	فصل اول : مروری بر تحلیل خوشه‌ای
۵	مقدمه
۶	۱.۱ نکاتی چند در تحلیل خوشه‌ای
۷	۲.۱ انواع روشهایی خوشه‌بندی
۸	۳.۱ روشهای خوشه‌بندی سلسله مراتبی
۸	۱.۳.۱ روشهای سلسله مراتبی تقسیمی
۹	۲.۳.۱ روشهای سلسله مراتبی انباشتی
۱۰	۴.۱ فاصله‌ها و مشابهتها
۱۰	۱.۴.۱ فاصله‌ها (عدم مشابهت)
۱۱	الف) فاصله‌ها برای داده‌های کمی
۱۲	ب) انواع فاصله‌ها برای داده‌های کیفی (جامعه‌های چند جمله‌ای)
۱۵	۲.۴.۱ مشابهتها
۱۵	۵.۱ الگوریتم خوشه‌بندی سلسله مراتبی انباشتی
۱۶	۶.۱ تحلیل خوشه‌ای به روش تک پیوندی
۲۰	۷.۱ تحلیل خوشه‌ای به روش پیوند کامل
۲۲	۸.۱ تحلیل خوشه‌ای به روش پیوند میانگین
۲۲	۹.۱ مشکلات تحلیل خوشه‌ای
۲۴	فصل دوم : نظریه‌گرافهای تصادفی
۲۵	۱.۲ تعاریف و مفاهیم نظریه‌گراف
۲۶	۱.۱.۲ همبندی گرافها
۲۶	۲.۱.۲ درختها
۲۷	۲.۲ مدل‌های متفاوت گرافهای تصادفی
۲۹	۳.۲ خاصیت گرافی
۳۱	۴.۲ ارتباط بین مدل‌های $G(n,p), G(n,M)$
۳۳	۵.۲ تابع آستانه‌ای
۳۳	۱.۵.۲ تابع آستانه‌ای برای مدل $G_M$
۳۴	۲.۵.۲ تابع آستانه‌ای برای مدل $G_p$
۳۴	۶.۲ مدل‌های دیگری از گرافهای تصادفی
۳۴	۱.۶.۲ مدل $G(n,K-out)$
۳۵	۲.۶.۲ مدل $G(n,p_1, \dots, p_K)$

۳۶	۷.۲ شکل کلی مدل‌های $G(n,p)$ و $G(n,M)$ .....
۳۶	۸.۲ فرآیند گراف تصادفی .....
۳۷	۹.۲ دنباله درجه‌ها در مدل $G(n,p)$ .....
۳۸	۱۰.۲ تکامل گرافهای تصادفی .....
۳۹	۱.۱۰.۲ مؤلفه‌های درختی .....
۴۲	۲.۱۰.۲ مؤلفه‌های دوری .....
۴۲	۳.۱۰.۲ مؤلفه‌های غول پیکر .....
۴۵	۱۱.۲ دسته‌ها در گراف $G_p$ .....
۴۶	۱۲.۲ تعداد مؤلفه‌ها در یک گراف $G_{n,M}$ .....
۴۷	۱۳.۲ مرتبه بزرگترین مؤلفه در یک گراف $G_{n,M}$ .....
۵۰	فصل سوم: تحلیل خوشه‌ای براساس گرافهای تصادفی .....
۵۱	مقدمه: .....
۵۳	۱.۳ تعریفها و نمادگذاریهای اصلی .....
۵۳	۱.۲.۳ درختواره‌نگارها .....
۵۵	۲.۲.۳ الگوریتم سلسله مراتبی انباشتی .....
۵۸	۳.۲.۳ معرفی آزمونهای فرض غیر قابل رده‌بندی بودن .....
۶۰	۳.۳ نتایج دقیق برای RSLIDها .....
۶۳	۱.۳.۳ سطوح افزازها .....
۶۵	۲.۳.۳ زمان بقای یک مجموعه تک عضوی .....
۶۸	۳.۳.۳ فاصله فرامتریک بین دو شیء .....
۷۳	۴.۳.۳ اندازه خوشه‌ها .....
۷۵	۴.۳ نتایج جانبی .....
۷۶	۱.۴.۳ رفتار تکاملی عام .....
۸۰	۲.۴.۳ سطوح افزاز .....
۸۲	۳.۴.۳ زمانهای بقای یک مجموعه تک عضوی .....
۸۳	۴.۴.۳ فواصل فرامتریک برای دو شیء .....
۸۷	۵.۴.۳ اندازه‌های خوشه‌ای .....
۸۸	۵.۳ مثالهایی از آزمون رده‌بندی .....
۸۹	۱.۵.۳ آزمون $H_0^{cont}$ در برابر $H_\mu^{cont}$ با استفاده از آخرین سطح .....
۹۳	۲.۵.۳ آزمون $H_0^{cont}$ در برابر $H_\mu^{cont}$ با استفاده از فاصله فرامتریک بین دو شیء .....
۹۶	فصل چهارم: رده‌بندی آبهای رودخانه کارون .....
۹۶	مقدمه .....



۹۹	۱.۴ ایستگاههای اندازه گیری
۱۰۰	۲.۴ قابلیت هدایت الکتریکی آب
۱۰۱	۱.۲.۴ محاسبه تقریبی کنداکتیویته آب
۱۰۱	۲.۲.۴ موارد استفاده از قابلیت هدایت الکتریکی
۱۰۲	۳.۲.۴ رده بندی آنها نسبت به قابلیت هدایت الکتریکی
۱۰۴	۳.۴ رده بندی آنها نسبت به سدیم
۱۰۶	۴.۴ رده بندی آبهای آبیاری
۱۰۹	۵.۴ تجزیه و تحلیل
۱۰۹	۱.۵.۴ تعریف یک فاصله (عدم تشابهت)
۱۱۰	۲.۵.۴ توزیع فراوانی فاصله ها
۱۱۱	۳.۵.۴ آزمون عدم رده بندی
۱۱۳	ضمیمه
۱۱۶	واژه نامه
۱۲۳	مراجع
۱۲۷	چکیده انگلیسی

## چکیده

این رساله درختواره‌نگارهای شاخص‌دار تصادفی تولید شده از الگوریتم‌های سلسله‌مراتبی 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 دنباله شاخصها، زمان بقای مجموعه‌های تک‌عضوی، فاصله

فرامتریک بین دو شیء مفروض یا اندازه‌های خوشه‌ای تعریف می‌شوند. برای یک درختواره شاخص‌دار

که به وسیله روش تک‌پیوندی 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 تک‌پیوندی

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 دهنده وجود نوعی از افراز در شیءها است، استفاده می‌شود. در پایان

نیز با استفاده از این آزمونهای جدید امکان رده‌بندی آبهای رودخانه کارون در استان خوزستان را مورد

بررسی و تجزیه و تحلیل قرار می‌دهیم.

پشگفتار

روشها و الگوریتمهای مختلفی برای رده‌بندی کردن یک مجموعه متناهی از مشاهدات وجود دارد. ورودیهای الگوریتمها می‌توانند عدم مشابهتها یا مقادیر یک مجموعه از متغیرهای اندازه‌گیری 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 تصمیم‌گیری، این مسئله مهم است که آیا خروجیهای حاصل از یک الگوریتم رده‌بندی، به یک خوشه‌بندی واقعی از داده‌ها اشاره می‌کند؟ ایده 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 آستانه مناسب بر پایه 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 [۳۱ و ۳۲]، [۴ و ۵ و ۶]، [۱۸ و ۱۹ و ۲۰] و [۲۴] و تعاریف متفاوت از توزیعهای احتمال مربوط به رده‌بندی تصادفی را در [۱۵]، [۳۰]، [۳۳]، [۱۹ و ۲۰] و [۶] پیدا کرد.

در این رساله الگوریتمهای سلسله‌مراتبی انباشتی را برای الگوریتمهای تک پیوندی [۱۳ و ۳۷]، پیوندی

کامل [۳۹] و پیوند میانگین در نظر گرفته ایم [۴۱]. این روشها یک درختواره نگار شاخص دار را از یک ماتریس که تفاوت دو به دوی اشیا را نشان می دهد، تولید می کنند. درختواره نگارهای شاخص دار به طور مستقل به وسیله هارتیگان<sup>(۱)</sup> [۳۳]، جاردین و سیبسون<sup>(۲)</sup> [۲۶] و جانسون<sup>(۳)</sup> [۲۸] تهیه شده اند.

در فصل اول این رساله به مفاهیم اساسی از تحلیل خوشه ای، انواع روشهای خوشه بندی و تعاریفی از فاصله (مشابهت) بین داده ها به عنوان ابزاری جهت رده بندی، همچنین بررسی سه روش تک پیوندی، پیوندی کامل و پیوند میانگین جهت خوشه بندی یک سری داده اشاره شده است.

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

در فصل دوم ابتدا تعاریفی از نظریه گرافهای تصادفی بیان می شود. سپس مدلهای متفاوت گرافهای تصادفی مانند  $G(n, P(\text{edge})=p)$  و  $G(n, M)$  تعریف و رابطه بین آنها مورد بررسی قرار می گیرد. در پایان نیز خصوصیتهای مهمی مانند وجود درختها، دورها، همبندی، همچنین زمان شکل گیری آنها در مدلهای ذکر شده در بالا در غالب قضیه هایی بیان شده است.

در فصل سوم تحلیل احتمالاتی الگوریتمها و بررسی رفتار یک الگوریتم با محاسبه توزیع احتمال خروجیها آن در حالت تصادفی بودن ورودیها صورت می پذیرد. در ابتدا تعاریفی از متغیرهای تصادفی مناسب، مربوط به آزمونهای آماری عدم رده بندی مشاهدات بیان می شود. در بخشهای بعدی (۳ و ۴) توزیعهای دقیق و مجانبی برای تمام متغیرهای تصادفی در حالتی از یک درختواره نگار تولید شده به وسیله روش پیوندی بر پایه تعاریفی از نظریه گراف [۷] است. در بخش آخر نیز آزمونهایی برای غیرقابل رده بندی بودن در مقابل فرضهای دیگر که نشان دهنده وجود افراز در شیءها می باشد، مطرح می شود.

۱) Hartigan

۲) Jardin and Sibson

۳) Johnson

در فصل چهارم، نحوه رده‌بندی آبهای رودخانه کارون در استان خوزستان با استفاده از دو فاکتور یا  $EC \times 10^6$  هدایت الکتریکی و SAR 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