بسم الله الرحمن الرحیم

دانشگاه تربیت معلم

دانشکده علوم ریاضی ومهندسی کامپیوتر

پایان نامه کارشناسی ارشد

رشته آمار

عنوان:

محاسبهٔ ماکسیمم آنتروپی دیریکله برای طرح ریزی داده‌ها

استاد راهنما:

دکتر عین ا... پاشا

نگارش:

نعمت مؤذن

مهر ۱۳۸۰

بسمه تعالی

آگهی دفاع از پایان‌نامه کارشناسی ارشد آمار

عنوان:

محاسبه ماکسیمم آنتروپی دیریکله برای طرح‌ریزی داده‌ها

استاد راهنما : آقای دکتر عین‌اله پاشا

داور خارجی : آقای دکتر غلامرضا محتشمی

داور داخلی : آقای دکتر علی‌اکبر رحیم‌زاده

دانشجو : **آقای نعمت مؤذن**

زمان : ساعت ۱۰ صبح روز شنبه مورخ ۸۰/۷/۱۴

مکان : دانشکدهٔ علوم ریاضی و مهندسی کامپیوتر دانشگاه تربیت معلم

خلاصه: اطلاع لگاریتمی که توسط شانون (۱۹۴۸) تعریف شد، کاربردهای زیادی در بخش‌های مختلف آمار پیدا کرده است. از جمله روش تشخیص توزیع نامعلوم مولد داده‌ها می‌باشد.

در این پایان‌نامه ابتدا به تعریف و بررسی بعضی خواص توابع اطلاع از قبیل آنتروپی، چگالی آنتروپی ماکسیمم، اطلاع تمییز ، قابلیت تمییز پذیری اطلاع (ID) و اطلاع متقابل می‌پردازیم. سپس مشخصهٔ آنتروپی ماکسیمم از خانواده پارامتری را با طرح پیشین دیریکله برای توزیع نامعلوم مولد داده‌ها ترکیب می‌کنیم که آن را روش آنتروپی ماکسیمم دیریکله (MED) می‌نامیم. روش یافتن توزیع‌های پیشین و پسین MED برای آنتروپی توزیع نامعلوم، پارامترهای گشتاورهای مدل تحت فرض، آنتروپی مدل ME و شاخص ID با استفاده از شبیه‌سازی «مونت کارلو» برای تشخیص توزیع نامعلوم مولد داده‌ها، از مباحثی هستند که در این پایان‌نامه بررسی می‌شوند.

در این پایان‌نامه مقالات زیر مورد استفادهٔ اصلی می‌باشد.

1. Soofi, E.S., N. Ebrahimi and M. Habibullah (1995), "Information Distinguishability with Application to Analysis of Failur Data", Journal of the American Statistical Assocciation, 90, 657-668.

2. Mazzuchi, T. A., Soofi, E.S. and Soyer, R. (2000) "Computation of Maximum Entropy Dirichlet for Modeling Lifetime Data". Computational Statistics and Data Analysis, 32, 361-378.

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

بسمه تعالی

دانشکدهٔ علوم ریاضی و مهندسی کامپیوتر

صورتجلسهٔ دفاع از پایان نامهٔ کارشناسی ارشد

جلسهٔ دفاع از پایان نامه آقای نعمت مؤذن دانشجوی دورهٔ کارشناسی ارشد رشتهٔ ریاضی شاخهٔ آمار تحت عنوان:

محاسبه ماکسیمم آنتروپی دیریکله برای طرح‌ریزی داده‌ها

در روز شنبه مورخهٔ ۸۰/۷/۱۴ در دانشکدهٔ علوم ریاضی و مهندسی کامپیوتر تشکیل گردید و نتیجه آزمون به شرح زیر تعیین می‌گردد. نمرهٔ این آزمون ۱۸٫۵ (هیجده و پنج) می‌باشد.

۱- عالی ☑
۲- بسیار خوب ☐
۳- خوب ☐
۴- قابل قبول ☐
۵- غیر قابل قبول ☐

استاد راهنما
دکتر عین‌اله پاسها

داور خارجی
دکتر غلامرضا محتشمی

داور داخلی
دکتر علی‌اکبر رحیم‌زاده

اسماعیل بابلیان
رئیس دانشکدهٔ علوم ریاضی و
مهندسی کامپیوتر

تهران، خیابان طالقانی بین بهار وشریعتی، پلاک ۵۹۹ کدپستی ۱۵۶۱۸ تلفن ۷۵۰۷۷۷۲ فاکس ۷۶۲۹۸۸

تقدیم به :

تقدیم به مادر عزیز و مهربانم
که در تمام مراحل تحصیل مشوق من بوده است.
تقدیم به برادران و خواهران بزرگوارم
که زحمات زیادی را در طول نگارش پایان نامه
به همراه همسران گرامیشان متحمل شدند.

تقدیر و سپاس

لازم است در اینجا از همه کسانی که در تدوین این پایان نامه مرا یاری نمودند سپاسگزاری کنم.از استاد گرانقدرم جناب آقای دکتر پاشا که تدوین این پایان نامه با راهنمایی های ایشان صورت گرفت تشکر وقدردانی می نمایم. همچنین از اساتید محترم جناب آقای دکتر رحیم زاده و جناب آقای دکتر محتشمی که قبول زحمت نموده و داوری این رساله را به عهده گرفتند کمال تشکر را دارم. در پایان از دوستان کارشناسی ارشد نیز بخاطر همراهیشان در این تحقیق نهایت قدردانی را می نمایم.

چکیده

اطلاع لگاریتمی که توسط شانون(۱۹٤۸) تعریف شد، کاربردهای زیادی در بخش‌های مختلف آمار پیدا کرده است. از جمله روش تشخیص توزیع نامعلوم مولد داده‌ها می‌باشد.

در این پایان‌نامه ابتدا به تعریف و بررسی بعضی خواص توابع اطلاع از قبیل آنتروپی، چگالی آنتروپی ماکسیمم، اطلاع تمییز ، قابلیت تمییز پذیری اطلاع (ID) و اطلاع متقابل می‌پردازیم. سپس مشخصه آنتروپی ماکسیمم از خانواده پارامتری ز با طرح پیشین دیریکله برای توزیع نامعلوم مولد داده‌ها ترکیب می‌کنیم که آن را روش آنتروپی ماکسیمم دیریکله (MED) می‌نامیم. روش یافتن توزیع‌های پیشین و پسین MED برای آنتروپی توزیع نامعلوم، پارامترهای گشتاوری مدل تحت فرض، آنتروپی مدل ME و شاخص ID با استفاده از شبیه سازی «مونت کارلو» برای تشخیص توزیع نامعلوم مولد داده‌ها، از مباحثی هستند که در این پایان نامه بررسی می‌شوند.

در این پایان‌نامه مقالات زیر مورد استفادهٔ اصلی می‌باشد.

1. Soofi, E.S. , N. Ebrahimi and M. Habibullah (1995), "Information Distinguishability with Application to Analysis of Failure Data", Journal of the American Statistical Association , 90, 657-668.

2. Mazzuchi, T. A, Soofi, E. S. and Soyer, R.(2000) " Computation of Maximum Entropy Dirichlet for Modeling Lifetime Data", Computational Statistics and Data Analysis, 32, 361-378.

فهرست

عنوان **صفحه**

پیشگفتار

هدف این تحقیق ارائه روشی بر پایه آنتروپی برای استنتاج بیزی در مورد اینکه یک توزیع پارامتری به عنوان یک مدل برای داده‌ها مناسب است یا خیر، می‌باشد. در زیر مروری اجمالی بر مباحث بررسی شده در این پایان نامه خواهیم داشت.

فصل اول: تعاریف و مقدمات آنتروپی

در این فصل ابتدا در بخش اول طرح متناهی و عدم حتمیت آن را تعریف کرده و سپس ارتباط میان آنتروپی و اطلاع را در بخش دوم شرح می‌دهیم. با استفاده از یک قضیه مهم اندازه‌ای برای عدم حتمیت به دست می‌آوریم این اندازه در واقع همان اندازه‌ای است که شانون (۱۹٤۸) نخستین بار معرفی نمود. در بخش سوم آنتروپی شرطی و اطلاع متقابل روی طرحها یا فضاهای متناهی بحث می‌شود. بخش چهارم نیز به آنتروپی متغیرهای تصادفی گسسته و پیوسته اختصاص یافته است که در آن با استفاده از قضایا، برخی خواص آنتروپی مورد بحث و بررسی قرار گرفته است.

فصل دوم: روش آنتروپی ماکسیمم و توابع توزیع دیریکله

در بخش نخست این فصل ابتدا به معرفی روش آنتروپیماکسیمم می‌پردازیم. سپس قضیه تابع چگالی ماکسیمم آنتروپی را با قیود گشتاوری داده شده بیان و اثبات می‌نماییم. در انتها جدول توابع چگالی آنتروپی ماکسیمم با قیود معلوم ارائه می‌شود. در بخش دوم به معرفی توابع توزیع دیریکله که در این پایان نامه استفاده شده است می‌پردازیم آنگاه به کاربرد آن در تعیین برآورد توزیع پیشین برای فرم‌های نامعلوم توابع توزیع خواهیم پرداخت.

فصل سوم: توابع اطلاع تمییز و خواص آنها

این فصل با هدف اندازه گیری تفاوت‌های کلی توزیع‌ها، روی تابع اطلاع «کالبک-لیبلر» بنا شده است. در بخش نخست به توضیح مقدمات استفاده از چگالی آنتروپی ماکسیمم در توابع اطلاع توسط محققان مختلف می‌پردازیم. در بخش بعدی برای مقایسه توزیع‌ها در Ω_θ ، اختلاف اطلاع بین دو چگالی f (نامعلوم) و f^* (آنتروپی ماکسیمم) مورد بررسی قرار گرفته است. تابع اطلاع کالبک - لیبلر به عنوان معروف‌ترین اندازه اطلاع نظری اختلاف بین توزیع‌ها شرح داده شده است. نامنفی بودن، جمع پذیری برای متغیرهای تصادفی مستقل از خواص این تابع می باشد. سپس شاخص ID توزیع‌ها، با استفاده از تابع اطلاع «کالبک- لیبلر» تعریف می‌شود که قابلیت تبعیض پذیری اطلاع را بیان می‌نماید و یک اندازه نرمال شده می‌باشد. به دست آوردن شاخص ID متناسب با حالتهای نرمال و نمایی از دیگر قسمتهای این بخش است. در بخش چهارم با معرفی آماره ID ، برآورد MLE این شاخص ارائه می‌شود. در بخش آخر نیز ارتباط بین اطلاع تمییز و اطلاع متقابل بیان می‌شود.

فصل چهارم: محاسبه ماکسیمم آنتروپی دیریکله برای طرح ریزی داده‌ها

در ابتدا کلاس گشتاوری مورد نظر و چگالی f^* در این کلاس معرفی شده سپس توابع اطلاع مورد نظر تعریف می‌شوند. در بخش دومآنتروپی کمی شده یک توزیع پیوسته f با افزاز نمودن خط حقیقی تعریف می‌شود. افراز مفیدی با استفاده از سنگفرشهای دیریکله ساخته می‌شود. برای تعیین توزیع F نامعلوم از یک طرح پیشین دیریکله استفاده می‌نماییم که یک توزیع کمی شده برای F به دست می آید. با استفاده از این توزیع، برآوردی برای آنتروپی محاسبه می‌شود. در بخش سوم،آنتروپی کمی شده، پارامترهای گشتاوری، مدل پارامتری و شاخص اطلاع ID با استفاده از پیشین دیریکله، برآورد شده و توزیع پیشین و پسین آنها با استفاده از یک الگوریتم شبیه‌سازی«مونت کارلو» به دست می‌آید.

مقدمه و تاریخچه

نظریه اطلاع برای اولین بار در سالهای ۴۸-۱۹۴۷ توسط "کلود شانون" مطرح شد.این نظریه در آغازبا بسیاری از مسائل ریاضی دشوار روبرو بود.طبیعی بود که شانون و اولین قوانینش که تمام هدف آنهامحاسبه نتایج سودمند بود، نمی‌توانست دقت کافی برای این مسائل مشکل ریاضی ارائه کند.

در نتیجه دربسیاری از این تحقیقات، دانشمندان به منظور سهولت اثباتها، مجبور می‌شدند با استدلال ازیک ماهیت غیرطبیعی یا مجموعه‌ای از موضوعات به طور مصنوعی (مانند، منابع، کانالها، کدها و غیره)استفاده کنند.بخش عمده ادبیات اولین سالهای نظریه اطلاع بناچار به محاسبات ناقص ریاضی که مخصوصاً آنرا برای‌ریاضی دانان بی نهایت مشکل می‌کرد، اختصاص یافت.

بعد یک کتاب درسی از نظریه اطلاع توسط "اس .گلدمن" منتشر شد که مثالهای رایجی را در آن بکاربرد.سپس مقالات زیادی نوشته شد که پایه‌های ریاضی و محاسباتی نظریه اطلاع را بنا نهادند.به عنوان‌اولین کارها باید از کار "مک میلان" نام برد که نظریه مفهوما ت اساسی منابع گسسته (منبع، کانال، کدوغیره) رابه عنوان اولین تعاریف دقیق ریاضی بکار برد.مهمترین نتیجه این کار برای اثبات قضیه مشهور،باید بکار گرفته می‌شد که در آن هر منبع آرگودیک گسسته ، ویژگی را که شانون به منابع نوع مارکف‌نسبت می‌داد دارا بود که این تقریباً همه محاسبات مجانبی نظریه اطلاع راارائه می‌داد.این وضع به تمام‌قسمتهای نظریه اطلاع گسسته اجازه می‌داد که بدون محدودیت، همانند نظریه شانون ،به منابع نوع‌مارکف ساخته شوند." مک میلان " در ادامه مقاله‌اش سعی نمود قضیه اساسی شانون را روی کانالهای بانوفه روی اصل دقیقی پایه ریزی کند که در انجام آن معلوم شد که اثبات خلاصه وار که توسط

شانون‌داده شده بود دارای وقفه و نقص زیادی است که حتی درحالت منابع مارکف هنوز باقی مانده بود.

سپس لازم است که به کار " فین اشتین " توجه شود که او مثل "مک میلان " فرض کرد که قضیه شانون‌روی کانالهای بانوفه، منتهای نظریه کلی اطلاع گسسته باشد و متعهد به دادن اثباتی دقیق از این نظریه شد.

با پذیرفتن عملیات ریاضی " مک میلان " با اجتناب از پیروی مسیر اصلی شانون، اثباتی را بنا نهاد.که درآن از یک عقیده، نتیجه بخش کاملاً جدید آشکار تحت عنوان " مجموعه قابل تمییز از دنباله‌ها "استفاده نمود.

هر چند " فین اشتین " اثبات آنرا برای مسائل و مسائلی که دارای بار عملی کمتری بودند بکار برد، که‌درآن سیگنالهای پی در پی منبع دوبدو مستقل هستند و حافظه کانال نیز صفر می‌باشد.در حالت کلی‌تر،فقط به طور خلاصه نشان داد که چطور می‌توان استدلال لازم را برای مستقل بودن انجام داد، متأسفانه ازلحاظ معنایی مشکلات زیادی باقی ماند.

کالبک ولیبلر (۱۹۵۱) اندازه فاصله بین دو چگالی مختلف را تعریف نمودند که نزدیکی و اطلاع تمییز دو چگالی را نشان می‌داد. این تابع آنتروپی یک چگالی نسبت به چگالی دیگر نیز می‌باشد.

جینس (۱۹۵۷) اصل آنتروپی ماکسیمم را بیان نمود. این اصل چگالی آنتروپی ماکسیمم را با قیود گشتاوری داده شده مشخص می‌کرد و به کاربردهای زیادی در علوم و زمینه‌های مختلف از جمله مکانیک آماری، نجوم، هواشناسی، اقتصاد، جغرافیا، تجارت بین المللی. بانکداری، توزیع جمعیتی، پزشکی و غیره دست یافت.

کالبک (۱۹۵۹) تابع اطلاع تمییز بین دو چگالی را با استفاده از اصل آنتروپی ماکسیمم بیان نمود که نزدیکی چگالی نامعلوم و چگالی آنتروپی ماکسیمم را نشان می‌دهد این پایان نامه براساس تابع اطلاع کالبک لیبلر (۱۹۵۹) بنا شده است و هدف آن تشریح توابع اطلاع و ارائه

روشی برای برآورد آنتروپی، توابع اطلاع، پارامترهای مدل و توزیعهای پیشین و پسین آنها با بکارگیری «ماکسیمم آنتروپی دیریکله» می‌باشد.