دانشگاه صنعتی اصفهان

دانشکده علوم ریاضی

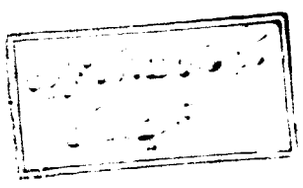
طبقه‌بندی داده‌ها با استفاده از مدلهای ترتیبی

پایان نامه کارشناسی ارشد آمار

آیتین سعادت‌ملی

استاد راهنما

دکتر ایوب ساعی



۱۳۷۹

دانشگاه صنعتی اصفهان

دانشکده علوم ریاضی

پایان نامه کارشناسی ارشد رشته آمار خانم آیتین سعادت‌ملی

تحت عنـــوان

طبقه‌بندی داده‌ها با استفاده از مدلهای ترتیبی

در تاریخ ۷۹/۷/۲۵ توسط کمیته تخصصی زیر مورد بررسی و تصویب نهائی قرار گرفت.

۱ ـ استاد راهنمای پایان‌نامه — دکتر ایوب ساعی

۲ ـ استاد مشاور پایان‌نامه — دکتر علی زینل همدانی

۳ ـ استاد داور ۱ — دکتر مجتبی گنجعلی

۴ ـ استاد داور ۲ — دکتر محمد صالحی

سرپرست تحصیلات تکمیلی دانشکده — دکتر امیر نادری

هذا من فضل ربی

حمد و سپاس خدای بزرگ و مهربان که همواره مرا مورد لطف و عنایت خویش قرار داده و توفیق انجام این پایان نامه را به من عطا فرمود.

بر خود لازم می‌دانم از کلیه عزیزان و سرورانی که مرا در انجام این پایان نامه یاری فرمودند کمال تشکر و قدردانی را بنمایم.

تشکر و سپاس ویژه از

جناب **دکتر ایوب ساعی** استاد راهنمای ارجمندم که با راهنمائیهای دلسوزانه و نیز جدیت و پشتکار بسیار، بنده را در ارائه این پایان نامه یاری فرمودند.

جناب **دکتر علی همدانی** استاد مشاور که کمال همکاری و مساعدت را فرمودند.

جناب **دکتر گنجعلی** استاد داور ارجمند که زحمت بازخوانی و داوری این پایان نامه را متقبل شدند.

جناب **دکتر محمد صالحی** که به عنوان استاد داور دانشکده در جلسه دفاع حضور یافته و تذکرات جالب و مفیدی را مطرح کردند و با پیشنهادات ظریف و بجا زمینه هر چه بهتر شدن پایان نامه را فراهم فرمودند.

جا دارد از کلیه اساتید دانشکده ریاضی دانشگاه صنعتی به ویژه **دکتر پارسیان** استاد گرانقدر و همچنین از کلیه اساتید گروه آمار دانشگاه اصفهان به ویژه **دکتر علامتساز** و به خصوص از خانواده خوب و فداکارم کمال تشکر و قدردانی را بنمایم.

تقدیم به

پدر و مادر فداکارم

خواهر و برادر عزیزم

فهرست مطالب

عنوان **صفحه**

چکیده

کمبل در سال ۱۹۹۱ یک روش شبیه سازی برای برازش مدل به داده‌های ترتیبی (Ordinal) انجام داد. نتایج شبیه سازی کمبل نشان داد که وقتی هدف اصلی طبقه بندی کردن داده‌ها باشد مدلهای ترتیبی هیچ مزیتی بر مدلهای غیر ترتیبی ندارد. کمبل در مقاله خود شبیه سازی را بدون در نظر گرفتن انواع مختلف داده‌ها و خواص آنها انجام داده است. در این تحقیق ما موضوع را با دیدگاه دیگری بررسی می‌کنیم. در این نوشتار با توجه به نوع گسسته یا پیوسته بودن متغیرهای کمکی مدلهای مختلفی را مورد بررسی قرار داده و انواع خطاهائی را که از این طبقه بندی حاصل می‌شود به دست می‌آوریم. همچنین نشان داده خواهد شد که مدلهای ترتیبی نسبت به مدلهای غیر ترتیبی بهتر عمل کنند. در آخر طبقه بندی کردن داده‌ها را با مدلهایی با اثرات آمیخته مورد بررسی قرار می‌دهیم.

فصل صفر

مقدمه:

در دهۀ 60 میلادی ابتدا واکرو دانکن [1] مدلهای مرتب شده را مورد توجه قرار دادند آنها مدلهای بخت متناسب را مورد توجه قرار دادند پس مکالگ [2] (۱۹۸۰) به بسط این مدلهاکه توسط واکر ودانکن مطرح شده بود پرداخت و آن را به فرم مدلهای لجستیک و مقدافرین ماکزیمال و مقدار فرین مینمال بسط داد و به بررسی آنها پرداخت.

در دهۀ ۷۰ایشنبیز و آوری [3] مدلهای تحلیل ممیزی را مورد بررسی قرار دادند که در سنوات بعد از آن به مدلهای تحلیل ممیزی نرمال با چندین گروه و سپس فرض با عدم همگنی، ماتریس واریانس کواریانس آن را بسط دادهاند که این مطلب اخیر از حوصلۀ این مقاله خارج است.

در سال ۱۹۸۱ فیلیپ و آندرسن [4] به بررسی رگرسیون مدلهای تحلیلی ممیزی پرداختند و قواعدی برای طبقه بندی آنها ارائه کردند.

در سال ۱۹۸۴ میلادی آندرسن [5] به بررسی رگرسیون متغیرهای طبقه بندی شدۀ مرتب شده پرداخت.

1- Walker fDuncan

2- Mccullagh

3- Eisenbers f Arery

4- Philips f Anderson

5- Anderson

بعد از آن ربشی و پولک [۱] به بررسی رگرسیون اینگونه متغیرها در صورت دوتائی بودند همت گماردند و آرمسترانگ و اسلوان [۲] مدلهای نسبت متناسب را در سال ۱۹۸۹ ابداع کردند و در آن مسائلی را که مربوط به اپیدمیولوژی بود بررسی کردند. در نهایت کمبل [۳] به بررسی خطاهای مربوط بـه مـدلهای مختلف پرداخت. البته او تنها خطای نوع اول که ترتیب را در نظر نمی‌گیرد مورد توجه قرار داد. در مقالهٔ حاضر سه خطای مختلف را که خطای نوع اول آن همان است که کمبل در نظر گرفته بود و ۲ خطای نوع دوم و سوم را مورد بررسی قرار داده است. خطای نوع دوم و سوم خطاهای مربوط به مـتغیرها را در حالت ترتیبی یعنی زمانی که ترتیب مهم است در نظر گرفته‌ایم. درفصل دوم به روشهای شبیه سازی در حالتی که متغیرهای کمکی دوتائی یا نرمال چند متغیره است پرداخته‌ایم. در فصل سوم به بررسی همین مدلها در حالتی که متغیرهای تصادفی نیز در مدل می‌باشد پرداخته‌ایم.

در فصل چهارم نتایج کامپیوتری و تحلیل آنها آمده است. تمام محاسبات با استفاده از نرم‌افزار SAS و زیر برنامهٔ PROC IML در آن انجام شده است.

1- Ashby f Pocok

2- Armstrong f sloan

3- cambell

فصل اول

معرفی مدلهائی برای طبقه بندی کردن داده‌ها

(۱-۱) مقدمه

متغیرهای پاسخ گسسته که بیش از دو طبقه داشته باشند اغلب ترتیبی می‌باشند. بدین معنی که پیشامدهای توصیف شده توسط اعداد طبقه ۱،۲،...،J را می‌توان همچون ترتیب در نظر گرفت. فرض کنید X یک بردار P بعدی از متغیرهای پیش بینی کننده باشد و y یک عدد مطابق با گروه عضویت باشد. فرض می‌کنیم که y , J مقدار ۱،۲،...،J را می‌گیرد و به علاوه گروههای y = J , Y = 1 را به ترتیب بهترین (نرمال) و بدترین (بیمارترین و زیان آورترین) خروجی‌ها می‌نامیم. ترتیب پذیری y به توزیع متغیرهای کمکی x منعکس می‌شود با میانگین آنها باید هم خطی چندگانه برای ترتیب پذیری مشاهدات داشته باشد.

ارتباط x,y ممکن است با مدلهای رگرسیون لجستیک بررسی شود. وقتیکه J=2 است y یک متغیر دو گروهی است؛ گروهها خوب / بد یا قبول / رد می‌باشد و مدل رگرسیون لجستیک دوتائی (MLR) قابل کاربردی است. در عمل J>2 است یک مدل رگرسیون لجستیک چند گروهی (MLR) خواهیم

داشت. یک مثال سیستمهای چند گروهی بد/ متوسط / خوب می‌باشد.

انگل [۱] در سال (۱۹۸۸) گروهبندیهای زیر را بکار برد.

۱- مدلهای عددی: این مدلها شامل تحلیل ممیزی نرمال (ND) و لجستیک چند گانه (ML) که در فصول بعدی شرح بیشتری دربارهٔ آن می‌دهیم می‌باشد.

۲- مدلهای تشخیصی: این مدلها حالتهائی که در آنها روی گروههائی که کارشناس تحلیل کرده است، شک و تردید وجود دارد را نشان می‌دهد. مثال این مدلها مدل لجستیک ترتیبی است.

۳- مدلهای بخت تجمعی: این مدلها فرض می‌کند که پاسخهای ترتیبی به یک متغیر پیوسته و غیر قابل مشاهده وابسته است و این متغیر پنهان دارای یک توزیع خاص در یکی از فواصل متصل بهم که بانقاط بریدگی معلوم از هم جدا شده است قرار می‌گیرند و با حرکت از یک فاصله به فاصلهٔ دیگر متغیر پاسخ تغییر می‌کند.

یک مثال متغیر پنهان مطابق با (سرعت) ملایم / متوسط / سریع می‌باشد. مدلهای بخت متناسب (PO) و نسبت دنباله‌ای (CR) دو زیر گروه این نوع مدلها می‌باشند.

در این ارتباط داریم: $\pi_{ij} = P\ (Y=j \mid X_i)$

بطوریکه در آن π_{ij} از طریق قضیهٔ بیز با عبارتهائی برای

$\pi_j = P\ (Y=j)$, $P\ (X_i \mid Y=j)$ بدست می‌آید.

گروه مرجع را ۱ = j در نظر می‌گیریم هدف از همهٔ فرایندهایی که در زیر توضیح می‌دهیم برآورد تابع چند جمله‌ای (۱-۱-۱) $\lambda_j(x_i) = \alpha_j + \beta_j' \times i$ می‌باشد.

(۲-۱) مدلهای عددی:

۱- مدلهای تحلیل ممیزی (ND):

این مدل توسط آوری [۲] و ایشینبیز [۳] (۱۹۷۵) پیشنهاد شد.

1- Engel

۲- Avery

۳- Eisenbeis

هدف از این تحلیل ممیزی چند متغیره این است که بر اساس اطلاعات قدرت جداسازی نمونه‌ها و یا صفات را بدست آورده و سپس برای مشاهدات جدید توانائی بهترین تخصیص را داشته باشیم. روش تحلیل ممیزی را یک روش جستجو گر وکشف کننده می‌نامیم.

اهداف استفاده از روش تحلیل ممیزی و طبقه بندی عبارتند از :

۱- تعیین تشخیص دهنده هائی که مقادیر عددی آنها برای مشاهدات چنان باشد که تا حد امکان جدا کردن آنها را ممکن سازد.

۲- مرتب نمودن مشاهدات و یا موضوعهای مورد مطالعه به چند گروه از قبل تعیین شده به طوریکه یک مشاهده را بتوان به صورات بهینه (یعنی حداقل خطا) به گروهی اختصاص داد.

فرمول بندی مسئله تحلیل ممیزی به این صورت است : فرض کنید J جمعیت متمایز $\pi_1, \ldots, \pi_J$ داریم. می‌خواهیم فردی را با مشاهده $x = (x_1, \ldots, x_p)'$ با P خصوصیت مختلف که فرد را مشخص می‌سازد به یکی از جمعیتهای $\pi_1, \ldots, \pi_J$ رده بندی کنیم.

وقتی گروه افراد را در نظر می‌گیریم اساساً فرض می‌کنیم که گروه به عنوان کل فقط به یکی از J جمعیت مفروض متعلق است. علاوه بر این فرض خواهیم کرد که هر یک از π_j ها را می‌توان بوسیله میانگینهای تابع توزیع F که به صورت یک بردار متغیر تصادفی $X = (X_1, \ldots, X_p)$ می‌باشد تعیین کرد به قسمی که مؤلفه‌های این بردار تصادفی اندازه‌های تصادفی برای P خصوصیت مختلف را نشان می‌دهند.

تمام فضای P بعدی مقادیر X را به E^p نشان می‌دهیم. اکنون می‌خواهیم برای تقیسم E^p به J ناحیه مجزای $R_1, \ldots, R_J$ قاعده‌ای تنظیم کنیم به گونه‌ای که اگر X در R_j قرار گیرد فرد را به گروه π_j اختصاص می دهیم، بدیهی است که در بکار بردن چنین قاعده‌ای ممکن است با اشتباه رده بندی کردن یک فرد به π_i برای $i \neq j$ وقتی واقعاً فرد به π_j متعلق است مرتکب خطا شویم.

چون حالت K جمعیت نرمال چند متغیره حالت کلی‌تر از ۲ جمعیت نرمال چند متغیره است ابتدا همان حالت $K = ۲$ را بررسی می‌کنیم.

فرض کنید که $y = j$ داده شده باشد، X دارای توزیع نرمال چند متغیره P بعدی با بردار میانگین μ_i و ماتریس واریانس کوواریانس Σ است. آنگاه تابع داده شد در (۱-۱-۱) را می‌خواهیم برآورد کنیم.

$$F_i(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} exp\left[\frac{-1}{2}(x-\mu_i)' \Sigma^{-1}(x-\mu_i)\right]$$

$$\frac{F_1(x)}{F_2(x)} = exp\left[\frac{-1}{2}(x-\mu_1)' \Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)' \Sigma^{-1}(x-\mu_2)\right]$$

$$= exp\left[\frac{-1}{2}(\mu_1-\mu_2)' \Sigma^{-1}(\mu_1+\mu_2) + (\mu_1-\mu_2)' \Sigma^{-1} x\right]$$

$$Ln\frac{F_1(x)}{F_2(x)} = (\mu_1-\mu_2)' \Sigma^{-1} x - \frac{1}{2}(\mu_1-\mu_2)' \Sigma^{-1}(\mu_1+\mu_2)$$

$$L(x) = (\mu_1-\mu_2)' \Sigma^{-1}\left[x - \frac{1}{2}(\mu_1+\mu_2)\right]$$

اما از طرفی

$$L(x) = \alpha_0 + \alpha_1 x$$

پس با مساوی قرار دادن دو عبارت $L(x)$ با هم داریم:

$$\alpha_0 = \frac{-1}{2}(\mu_1-\mu_2)' \Sigma^{-1}(\mu_1+\mu_2)$$

$$\alpha = \alpha_1 = (\mu_1-\mu_2)' \Sigma^{-1}$$

که عبارت اخیر یک بردار سطری می‌باشد.

از اینرو اگر

$$\alpha_0 + \alpha_1 x = (\mu_1-\mu_2)' \Sigma^{-1} x - \frac{1}{2}(\mu_1-\mu_2)' \Sigma(\mu_1+\mu_2)$$

$$\Rightarrow \hat{\alpha}_0 = \frac{-1}{2}(\mu_1-\mu_2)' \Sigma^{-1}(\mu_1+\mu_2)' \qquad (۱-۲-۱)$$

$$\hat{\alpha}_1 = (\mu_1-\mu_2)' \Sigma^{-1} \qquad (۱-۲-۲)$$