تحل ل مم ز ا استفاده از م خته‌ها نرمال

قاسم رکابدار[1] رح م چ نی‌پرداز[2]

[1] دانشگاه زاد اسلام واحد بادان خرمشهر

[2] دانشگاه شه د چمران گروه مار

چک ده تحل ل مم زی خ ی ف شر[1] (LDA) ا زار مهمی رای رده ندی ن چندگروه می اشد هنگام که کلاسها دارای توز نرمال چند متغ ره ا ماتر س کوار انس مشترک اشند بق لم ز من پ رسن LDA ه نه وده و معادل ا قاعده رده ندی ماکز مم درستنما ی است LDA دارای معا بی است از جمله فرض نرمال ودن کلاسها اغل ندرت پ ش می د و کرانهای تصم م خ ی کلاسها را ه قدر کافی از هم جدا نمی‌کنند رای تعم م LDA می‌توان فرض کرد که کلاسها از ز ر کلاسها ی تشک ل شده‌اند که دارای توز نرمال ا م انگ نهای متفاوت و ماتر س کوار انس مشترک هستند مم زی در ا ن حالت م خته[2] (MDA) می اشد و کرانهای تصم م ن کلاسها غ ر خ ی و پ چ ده وده و رورد پارامترها تنها از ر ق الگور تم EM امکان‌پذ ر است

کاهش ا عاد در مم زی کی از اهداف مهمی است که ه وس له مم زی خ ی ف شر امکان‌پذ ر می‌گردد و وس له متغ رهای مم زی ف شر می‌توان کلاسها را در نمودارهای دو عدی مشاهده نمود مم زی م خته ن ز ا ن خاص ت را در ردارد و کاهش ا عاد داده‌ها در ف ای رازش افته توس م انگ ن ز ر کلاسها امکان‌پذ ر است در ا ن م العه مم زی وس له م خته‌های نرمال توسعه داده می‌شود و سپس داده‌های گزارش توسعه انسانی در سال را رای مقا سه روشهای مم زی رای رده ندی کشورها ه سه کلاس کشورها ا توسعه انسانی الا کشورها ا توسعه انسانی متوس و کشورها توسعه انسانی پا ن در ن ر گرفته شدند و نشان داده شد که مم زی م خته دارای نرخ خ ای کمتری نسبت ه د گر روشها رای رده ندی کشورها است

واژه‌ها کل د رده ندی تحل ل خوشه‌ای الگور تم EM تحل ل مم زی انع اف پذ ر شاخص توسعه انسانی

مقدمه

رده ندی دو ا چند کلاس ا مم زی از مو وعات مهم در مار چند متغ ره است که دارای کار ردهای مختلف در حوزه‌های گوناگون است در حالت کلی مم زی اختصاص ک ا چند مشاهده x ا

1) Linear Discriminant Analysis 2) Mixture Discriminant Analysis

کلاس نامعلوم ه جوامعی معلوم $G_1, G_2, ..., G_J$ است رای ا ن کار فرض می‌شود مجموعه راهنما $\{x_i, g_i\}_{i=1}^{n}$ است که در ن $x_i \in \Re^p$ موجود وg_i مشخص کننده کلاس نمونه iام است ا استفاده از مجموعه راهنما پارامترها در مدل رورد و سپس ا استفاده از پارامترهای رورد شده هدف دست وردن ک قاعده رای مم زی صورت ک تا جهت پ ش نی مشاهده جد د x_0 ه کی از کلاسها است

روش سنتی و کلاس ک ماری مم زی خ ی و رگرس ون لجست ک است ا ن روش‌ها تا حدود ز ادی توسعه افته و معمولا در متون ماری موجود هستند ه عنوان مثال مارد ا و د گران و ندرسن رای مم زی خ ی و مم زی لجست ک مراج خو ی محسو می‌شوند هرگاه جوام نرمال اشند تا مم زی خ ی ز نرمال و نا را ن خ ای مم زی ز راحتی قا ل محاسبه خواهد ود روش درجه دوم[3] در مم زی دل ل پ چ ده ودن توز مم زی وس له ماردانان هنوز در حال توسعه است و روش‌های مم زی غ ر خ ی از جمله درجه دوم ناپارامتری و . . . را مورد رسی قرار داده است

متاسفانه فرض نرمال ودن کلاسها ندرت پ ش می د در چ ن صورتی تا مم زی داده‌ها را صورت موثری از هم جدا نمی‌کند و خ ای مم زی افزا ش خواهد افت ک ر ق رای حل ا ن مساله فرض تعم م LDA است ه ا نکه فرض شود کلاسها شامل ز رکلاسها ی نرمال‌اند عنی کلاسها دارای توز م خته‌اند و نا را ن مم زی کار رفته ز م خته خواهد ود ا ن مو و در نوشتجات ماری وس له مک لاچلان و تکست و د گران پ شنهاد شده است هاستی و ت بش رانی ا ن م ل را گسترش داده‌اند نها فرض کرده‌اند هر کلاس از ز ر کلاسها ی ا م انگ ن متفاوت و ماتر س کوار انس درون و ن کلاس‌های مشترک هستند LDA فق رای رده ندی استفاده نمی‌شود لکه ا زار سودمندی رای کاهش ا عاد داده‌هاست ا LDA می‌توان چند ن کلاس از داده‌ها را در غال نمودارهای دو عدی نما ش داد گونه‌ای که شتر ن جداسازی ن م انگ ن کلاسها را داشته اشند MDA ز ا ن خاص ت را دارد و می‌توان کاهش ا عاد را روی م انگ ن کلاسها ز رکلاسها مشاهده نمود در LDA کاهش ا عاد در حالتی که مم زی ن دو کلاس اشد امکان‌پذ ر ست اما در MDA ا ن امر امکان‌پذ ر است

در ا ن مقاله تحل ل مم زی م خته رسی و الگور تم EM رای رورد پارامترها و مم زی پ شنهاد شده است ا ن الگور تم وس له هاستی و ت بش رانی پ شنهاد و وس له نو سندگان مقاله از باق و سپس در داده‌های توسعه انسانی کارگرفته شده است و در خش دوم مقاله‌های مم زی خ ی مرور و خش سوم مقاله توسعه LDA ه MDA را شامل می‌شود خش چهارم ه از باق الگور تم EM رای استفاده در مم زی MDA می‌پردازد و در خش پنجم داده‌های توسعه انسانی را رای مقا سه ن MDA ا LDA و QDA در ز ر گرفته شده‌اند سرانجام خش‌های نها ی مقاله ک نت جه‌گ ری در خصوص MDA خواهد ود

3) Quadratic Discriminant Analysis

تحل ل مم ز خ LDA

فرض کن د در مساله مم زی مجموعه راهنما $\{x_i, g_i\}_{i=1}^{n}$ در ا نصورت چگالی شر ی X در کلاس jام ه صورت ز ر خواهد ود

$$p_j(x) = \phi(\mu_j, \Sigma) = P(X = x \mid G = j)$$

$$= \mid \ \pi\Sigma \mid^{-\frac{1}{2}} exp\left\{ --D(x_i, \mu_j)\right\} \quad (\ )$$

در ا نصورت لگار تم درستنما ی مشاهدات ه صورت ز ر خواهد ود

$$\ell(\mu_j, \Sigma) = -\sum_{j=1}^{J}\sum_{g_i=j} D(x_i, \mu_j) - nlog \mid \Sigma \mid \quad (\ )$$

در ا نجا $D(x,\mu) = (x-\mu)'\Sigma^{-1}(x-\mu)$ فاصله ماهالانو س ن x و μ و $\Sigma_{g_i=j}$ ه معنی مجمو مشاهدات مجموعه راهنما در کلاس jام است گ ر م احتمال پ ش ن کلاس jام صورت $_j =$ $P(G=j) =$ اشد که معمولا از قبل معلوم است و ا از مجموعه راهنما رورد می‌شود اگر ه چ دل لی ر ترج ح کلاسها ر کد گر نباشد $_J = \frac{1}{J}$ $= \ldots =$ $_2 =$ $_1$ در ن ر گرفته می‌شود سپس مشاهده x_0 ه کلاس jام رده ندی می‌شود $C(x_0) = j$ اگر احتمال پس ن ا نکه از کلاس jام اشد در م ان ق ه کلاسها ماکز مم شود

$$P(G = j \mid X = x_0) = max_\ell P(G = \ell \mid X = x_0) \quad (\ )$$

هر گاه $P(G=j) = \frac{1}{J}$ اشد ا ن روش ا روش مم زی ماکز مم درستنما ی معادل می‌شود ا توجه ه اگر کلاسها دارای توز نرمال ا ماتر س کوار انس مشترک اشند چگالی پس ن کلاس jام ه صورت ز ر خواهد ود

$$P(G = j \mid X = x) = \frac{_j\phi(\mu_j, \Sigma)}{\sum_{l=1}^{J} \ _l\phi(\mu_l, \Sigma)}$$

$$\propto exp\left\{x'\Sigma^{-1}\mu_j - -\mu_j'\Sigma^{-1}\mu_j + \log \ _j\right\}$$

$$\propto exp\left\{x'\beta_j + \alpha_j\right\}$$

توجه کن د ثا ت ودن مخرج کسر نسبت ه j تاثری در ماکز مم کردن کلاسها ندارد کران تصم م ن دو کلاس مانند i و jام ه صورت مجموعه‌ای از نقا تعر ف می‌شود که دارای احتمالات پس ن مساوی هستند عنی

$$b_{i,j} = \{x \in \Re^p; P(G = j \mid X = x) = P(G = i \mid X = x)\}$$

و نا را ن کران تصم م   ن دو کلاس صورت ز ر است

$$(\beta_j - \beta_i)'x + (\alpha_j - \alpha_i) =$$

که تا عی خ ی از x است   تا   مم زی رای کلاس jام ه صورت ز ر تعر ف می‌شود

$$\delta_j(x) = -(x'\beta_j + \alpha_j) \qquad (\ )$$

و قاعده رده ندی انتسا $x_\circ$ ه کلاس jام است اگر را   ه ز ر رقرار اشد

$$\delta_j(x) = -min_\ell(x'\beta_\ell + \alpha_\ell)$$

 ا د توجه شود که در حالت مساوی ودن احتمال پ ش ن کلاسها مشاهده جد د $x_\circ$ ه کلاسی رده ندی می‌شود که کمتر ن فاصله ماهالانو س را ا مرکز ن داشته اشد α_ℓ و β_ℓ ا استفاده از مجموعه راهنما رورد می‌شوند نا را ن ماکز مم   صورت ز ر حاصل می‌شوند

$$\hat{\mu}_j = \frac{}{n_j}\sum_{g_i=j} x_i \qquad \hat{\Sigma} = \frac{}{n}\sum_{g_i=j}^{J}\sum_{g_i=j}(x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)' \qquad (\ )$$

لازم ذکر است اگر کلاسهای دارای ماتر س کوار انس مختلف اشند عنی

$$p_j(x) = \phi(\mu_j, \Sigma_j)$$

کران تصم م   ن دو کلاس i و j غ ر خ ی ه صورت تا   درجه دوم از x خواهد ود در ا ن حالت تعداد پارامترها ی که ا د رورد شوند $JP + \frac{P(P+1)}{2}$ خواهد ود چون اندازه نمونه‌های مجموعه راهنما ثا ت است نا را ن رورد پارامترها در مم زی درجه دوم ن رومند ن ست و رای حالت $P >$ پ شنهاد نمی‌شود

۱ کاهش ابعاد در LDA

کاهش ا عاد در مم زی را می‌توان ا تجز ه ف شر   دست ورد در روش پ شنهادی ف شر هدف دست وردن $\alpha \in \Re^p$ گونه‌ای که کسر ز ر ماکز مم شود

$$\frac{\alpha' B \alpha}{\alpha' W \alpha}$$

در ا نجا B ماتر س کوار انس   ن گروهها و W ماتر س کوار انس درون گروهی است   هتر ن جوا   رای α ردار و ژه م ا ق ا زرگتر ن مقدار و ژه ماتر س $W^{-1}B$ است در حالت کلی ا ن ماتر س حداکثر دارای $min(J - \ , p)$ مقدار و ژه غ ر صفر است نا را ن متغ رهای مم زی

ف شر ه صورت $y_k = \alpha'_k x$ و $k = \ \ , \ \ , ..., K$ می اشند که $K \leq min(J - \ \ , p)$ عنی عملا از K مم ز اول ف شر در LDA ه عنوان متغ ر مم زی استفاده می‌شود دو متغ ر مم زی اول ف شر شتر ن جداسازی را ن م انگ ن کلاسها دارند و رای نما ش کلاسها در غال نمودار استفاده می‌شوند   در مم زی خ ی ف شر مشاهده جد د $x_\circ$ ه کلاس jام رده ندی می‌شود اگر فاصله اقل دسی ز ر کمتر ن در م ان همه کلاسها اشد

$$\delta_j^{\text{۲}}(x_\circ) = \sum_{k=\text{۱}}^{K} (y_k(x_\circ) - \bar{y}_k^j)^{\text{۲}}$$

که $\bar{y}_k^j$ م انگ ن متغ ر مم زی kام در کلاس jام است   در حالت نرمال ودن کلاسها هاستی و ت بش رانی   نشان داده‌اند که رده ندی در ف ای مم زی خ ی ف شر معادل ا LDA است همچن ن هاستی و د گران   نشان دادن که ف ای مم زی خ ی ف شر (LDA) معادل ا رگرس ون خ ی چند متغ ره است که از مقاد ر به نه رای نشان دادن کلاسها استفاده شده است در حالت کلی LDA را می‌توان صورت دنباله‌ای از رگرس ونهای چندگانه خ ی $\eta_k(X) = X'\beta_k$ و $k = \ \ , ..., K$ نما ش داد ا ن شکل از مم زی پس از تع ن مقاد ر به نه $\theta_\text{۱}, ..., \theta_K$ و انتسا نها ه کد کلاسها دست ورده می‌شود   روش کار صورت ز ر است فرض شود که Y ماتر سی است که اگر x_i متعلق ه کلاس jام اشد نگاه در رد ف iام و ستون jام مقدار ن ک است و ق ه ستونها در رد ف iام مقدار صفر را دارند

 رگرس ون چند متغ ره خ ی   رگرس ون خ ی چند متغ ره از Y در مقا ل X ماتر س پ ش نی‌گرها   رازش می‌شود گ ر م $\hat{Y}$ ماتر س مقاد ر رازش شده اشد و $\eta(X)$ ردار توا   رگرس ونی اشد

 مقاد ر به نه K تا از زرگتر ن ردارهای و ژه $Y'\hat{Y}$ ا شر $\Theta' D_\Pi \Theta = I_K$ که $D_\Pi = \frac{Y'Y}{n}$ ماتر س ق ری از احتمال پ ش ن کلاسهاست فرض شود $\Theta J \times K$ ماتر س مقاد ر به نه   ر ورد شده اشد

 ا استفاده از ماتر س مقاد ر به نه در مرحله   و ردار توا   رگرس ونی در مرحله $\eta(X) \longleftarrow \Theta'\eta(X)$

 در ا ن حالت مشاهده جد د $x_\circ$ ه کلاس jام رده ندی می‌شود اگر فاصله اقل دسی ز ر در م ان همه کلاسها کمتر ن اشد

$$\delta_j^{\text{۲}}(x_\circ) = \sum_{k=\text{۱}}^{K} w_k(\eta_k(x_\circ) - \bar{\eta}_k^j(x)^{\text{۲}}$$

که $\bar{\eta}_k^j(x)$ م انگ ن مقاد ر رازش شده تا   رگرس ونی $\eta_k(x)$ درکلاس jام است و $w_k = \frac{\text{۱}}{r_k^{\text{۲}}(\text{۱} - r_k^{\text{۲}})}$ وزنها ی هستند که ف ای رازش شده رگرس ونی را ه ف ای مم زی ف شر تبد ل می‌کنند و r_k kام مقدار و ژه‌ای است که در مرحله   الگور تم محاسبه می‌شود   اگر در مرحله   از الگور تم

از رگرسیون ناپارامتری استفاده شود در این صورت مرز غیر خطی خواهد بود هاستی و دیگران هر روشی را که از یک پیش‌پردازش رگرسیونی برای مسایل مرزی استفاده شود را مرزی انعطاف‌پذیر[4] (FDA) نامیده‌اند بنابراین اگر پیش‌پردازش رگرسیون خطی باشد مرزی نیز خطی خواهد بود

تعمیم LDA با استفاده از مخلوط‌ها نرمال

فرض کنید در بخش قبلی تابع چگالی در هر کلاس مخلوط نرمال باشد که در آن هر مولفه دارای میانگین مربوط به خود و ماتریس کواریانس مشترک با بقیه مولفه‌ها باشد بنابراین در کلاس jام به صورت زیر خواهد بود

$$p_j(x) = \sum_{r=1}^{R_j} \pi_{jr}\phi(x_i, \mu_{jr}, \Sigma)$$

$$= | \pi\Sigma |^{\frac{-1}{2}} exp\left\{--D(\mu_{jr}, \Sigma)\right\} \qquad ( )$$

در اینجا R_j تعداد مولفه‌های کلاس jام و π_{jr} نسبتهای مخلوطگی هر یک از مولفه‌های نرمال است که $\sum_{r=1}^{R_j} \hat{\pi}_{jr} =$ می باشد در مدل مرزی مخلوطه احتمال پسین کلاس jام در صورتی که x مشاهده شده باشد به صورت زیر است

$$P(G = j \mid X = x) = \frac{\sum_{r=1}^{R_j} \pi_{jr}\phi(x_i, \mu_{jr}, \Sigma)}{\sum_{j=1}^{J}\sum_{r=1}^{R_j} \pi_{jr}\phi(x_i, \mu_{jr}, \Sigma)} \qquad ( )$$

لگاریتم درستنمایی شرطی در اینجا صورت زیر است

$$\ell_m(\mu_{jr}, \Sigma, \pi jr) \propto \sum_{j=J}^{J} log(\sum_{r=1}^{R_j} \pi_{jr} exp\left\{--D(\mu_{jr}, \Sigma)\right\}) - \frac{n}{}log \mid \Sigma \mid \quad ( )$$

در MDA مشاهده جدید x_0 به کلاس jام نسبت داده می‌شود $C(x_0) = j$ اگر را به در میان تمام کلاسهای دیگر ماکزیمم شود

$$P(G = j \mid X = x_0) = max_\ell P(G = \ell \mid X = x_0)$$

4) Flexible Discriminant Analysis

در ا نجا پارامترهای مدل $R_j, \pi_j r, \mu_j r, \Sigma$ که $j = \ , ..., J, \quad r = \ , ..., R_j$ می اشد ا ن پارامترها از ر ق مجموعه راهنما ا ماکز مم کردن را ه دست می ند د هی است که در MDA ماکز مم کردن را ه از ر ق مشتق‌گ ری امکان‌پذ ر ن ست نا را ن ا د از ر ق روشهای عددی ماکز مم شود

روورد پارامتر مدل مم ز م خته

دمپستر و د گران رای رورد پارامترها م رح کرده‌اند که در هنگام ناکامل ودن مشاهدات کار رد دارد الگور تم EM روشی تکراری تا رس دن ه همگرا ی در لگار تم درستنما ی داده‌هاست که در هر تکرار مرحله E ام د ر ا ی و مرحله M رورد ماکز مم درستنما ی وجود دارد در مدل مم زی م خته ور دق ق مشخص ن ست که هر نمونه در مجموعه راهنما ه کدام ز ر کلاس تعلق دارد نا را ن رای ماکز مم کردن ا د از الگور تم عددی EM استفاده نمود

مرحله E

مرحله E از الگور تم EM شامل ام د ر ا ی لگار تم درستنما ی داده‌های کامل است ه شر ا نکه مقدار مشاهده شده x و مقدار صح ح پارامترها در هر تکرار در اخت ار اشد در داده‌های م خته نرمال می‌توان مشاهدات را ه صورت داده‌های ناکامل در ن ر گرفت که مقاد ر گمشده در ن کد ز ر کلاس هر مولفه می اشد در ا نجا ا فرض ا نکه پارامترهای ا تدا ی مشخص هستند احتمال ا وزن ا نکه مشاهده x_i متعلق ه ز ر کلاس rام از کلاس jام اشد رای همه نمونه‌های مجموعه راهنما در کلاس jام محاسبه می‌شود که رای اثبات می‌توان ه مک لاچلان مراجعه کرد

$$P(C_{jr}(x_i) \mid x_i, j) = \frac{\pi_{jr}\phi(x_i, \mu_{jr}, \Sigma)}{\sum_{l=۱}^{R_j} \pi_{jl}\phi(x_i \mu_{jl}, \Sigma)} \qquad (\ )$$

مرحله M

 ا استفاده از وزنهای احتمالات رورد شده در مرحله E را ه روردهای ماکز مم درستنما ی رای پارامترهای هر مولفه نرمال در هر کلاس محاسبه می‌شوند

$$\hat{\pi}_{jr} \propto \sum_{g_i = j} P(C_{jr}(x_i) \mid x_i, j) \qquad \sum_{r=۱}^{R_j} \hat{\pi}_{jr} = \qquad (\ )$$

$$\hat{\mu}_{jr} = \frac{\sum_{g_i=j} P(C_{jr}(x_i) \mid x_i, j) x_i}{\sum_{g_i=j} P(C_{jr}(x_i) \mid x_i, j)} \qquad (\ )$$

$$\hat{\Sigma} = \frac{}{n} \sum_{j=١}^{J} \sum_{g_i=j} \sum_{r=١}^{R_j} P(C_{jr}(x_i) \mid x_i, j)(x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)' \qquad (\quad)$$

هما‌ن ور که از مقا سه را ه ا را ه‌های و د ده می‌شود رورده‌های ماکز مم درستنما ی مانند حالت نرمال کامل هستند ا ا ن تفاوت که وزنهای $P(C_{jr}(x_i) \mid x_i, j)$ جا گز ن تا نشانگر کلاسها شده‌اند ا توجه ه ا نکه وزنهای $P(C_{jr}(x_i) \mid x_i, j)$ خود تا عی از μ_{jr} و Σ هستند را ه‌های تا در را ه جا گز ن شده تا وزنهای عدی محاسبه شوند ا ن مراحل نقدر تکرار می‌شوند تا همگرا ی در لگار تم درستنما ی شر ی روی دهد یکی از خواص مناسـ الگور تم EM ا ن است که در هر تکرار لگار تم درستنما ی داده‌ها نسبت ه تکرار قبلی افزا ش می ا د نا را ن اگر در مرحله‌ای را ه نسبت ه تکرار قبل از ن تغ ر کمی داشته اشد الگور تم متوقف می‌شود و رورده‌های نها ی دست ورده می‌شوند

تکرارهای الگور تم EM ه اندازه خوشه‌ها R_j و رورده‌های ا تدا ی رای پارامترهای μ_{jr} و Σ و وزنهای $P(C_{jr}(x_i) \mid x_i, j)$ دارند رای ا ن منظ ور می‌توان از الگور تم خوشه ندی م انگ ن Kام[5] ک مقدار ثا ت رای خوشه‌ها تعداد ز ر کلاسها تع ن کرد و سپس از ا ن الگور تم رای رورد مراکز ز ر کلاسها μ_{jr} در هر کلاس استفاده شود سپس رای همه مشاهدات مجموعه راهنما درکلاس jام $P(C_{jr}(x_i) \mid x_i, j)$ را ر ا ک است اگرμ_{jr} نزد کتر ن مرکز و در غ ر ا نصورت صفر است پس از رورد پارامترها مشاهده جد د x م ا ق ا را ه ه خوشه مورد ن ر نسبت داده می شود در MDA ا مساوی قرار دادن احتمال پس ن کلاسها ه تا عی پ چ ده و غ ر خ ی خواه م رس د نا را ن در MDA حتی ا فرض مساوی ودن ماتر س کوار انس کلاسها ز ر کلاسها کرانهای تصم م ن کلاسها غ رخ ی است که ا ن کی از تعم مهای مهمی است که هنگام فرض نرمال م خت ه رای کلاسها در LDA دست می د

فرض مساوی ودن ماتر س کوار انس ز ر کلاسها ا ن امکان را وجود می ورد تا کاهش ا عاد داده‌ها در MDA دست ورده شود ه عبارت د گر می‌توان فرض نمود که ا نسخه وزنی LDA رو رو هست م که ز ر کلاسها جا گز ن کلاسها شده‌اند اگر $R = \sum_j^J R_j$ تعداد کل ز ر کلاسها اشد که هر کدام از مشاهدات ا وزن خود ه هر ک از ز ر کلاسها تعلق می‌گ رد در ا ن حالت کلاس jام شامل $n_j R_j$ مشاهده است نا را ن مجموعه راهنما در ا ن حالت مجموعه‌ای وزنی و افزوده شده ه تعداد $n' = \sum_j^J n_j R_j$ از مشاهدات است دو اره از الگور تم EM کمک گرفته می‌شود و پس از محاسبه وزنها در مرحله E در مرحله M مس له LDA وزنی حل می‌شود نرم افزار مورد استفاده ا ن امکان را می‌دهد که در مرحله M از مقاد ر ه نه رای محاسبه متغ رهای مم زی استفاده شود اگر $Z_{n \times R}$ ماتر س پاسخ اهری اشد که رد فهای ا ن ماتر س وزن مشاهدات است گونه‌ای که اگر مشاهده iام متعلق ه کلاس ام اشد jام اشد نگاه ز ر کلاسهای مر و ه کلاس jام ا وزن ا ن مشاهده پر می‌شوند و ق ه ز ر کلاسها در رد ف iام مقدار صفر را می‌گ رند ا د توجه داشت که مجمو وزنها را ر n است $n = \sum_{j=١}^{J} \sum_{g_i=j} \sum_{r=١}^{R_j} P(C_{jr}(x_i) \mid x_i, j)$

5) Kmeans

جدول نرخ خطای رده‌بندی نادرست برای داده‌های توسعه انسانی

روش	مجموعه راهنما	مجموعه آزمون
LDA		
QDA		
MDA		

پس از محاسبه ماتریس ماهری Z با استفاده از مراحل گفته شده در بخش مدل کاهش یافته MDA در هر تکرار محاسبه می‌شود در LDA کاهش ابعاد داده‌ها توسط تعداد کلاسها محدود شده است بنابراین در حالتی که مرزی بین دو کلاس باشد $J =$ کاهش ابعاد امکان‌پذیر نیست اما در MDA چون زیرکلاسها جایگزین کلاسها می‌شوند بنابراین همواره متغیرهای مرزی قابل محاسبه بوده و کاهش ابعاد امکان‌پذیر است

مثال کاربرد رده‌بندی کشورها

از سال سازمان ملل متحد هر ساله گزارشی تحت عنوان گزارش توسعه انسانی را منتشر می‌کند در این گزارش برای رده‌بندی کشورها از نظر توسعه انسانی تنها متغیرهای زیر در نظر گرفته می‌شود

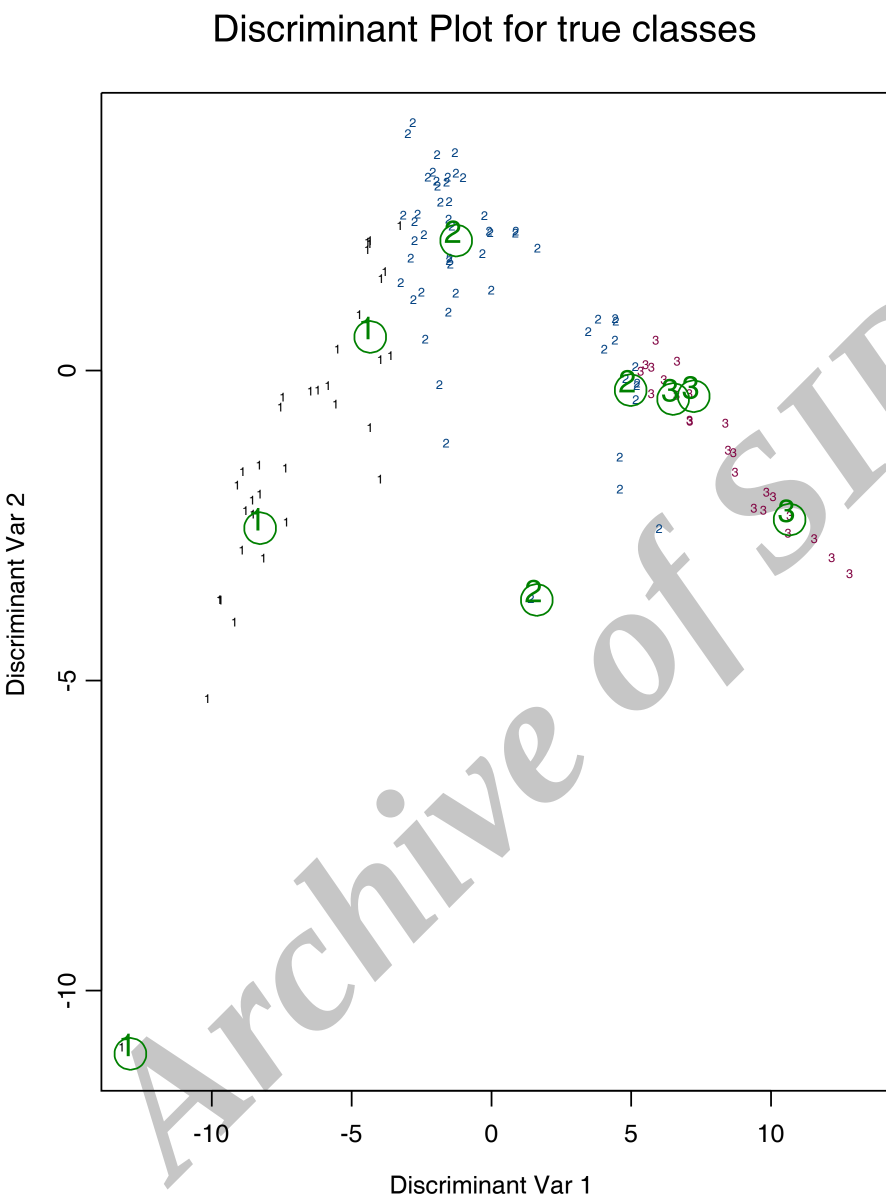
x_1 امید به زندگی در بدو تولد
x_2 نرخ باسوادی بزرگسالان
x_3 نرخ رشد نام نویسی در دوره‌های تحصیلی
x_4: سرانه محصول ناخالص داخلی

جمع‌آوری اطلاعات معمولا با استفاده از اطلاعات رسمی منتشره کشورها و در صورت غیر قابل اعتماد بودن اطلاعات وسیله برنامه عمران سازمان ملل برآورد می‌شوند در این گزارش از داده‌های بدست آمده شاخص توسعه انسانی[6] HDI محاسبه می‌شود و با استفاده از آن کشورها به سه رده کشورها با توسعه انسانی بالا کشورها با توسعه انسانی متوسط و کشورها با توسعه انسانی پایین رده‌بندی می‌شوند

۱ رده‌بندی کشورها با تحلیل ممیز

در گزارش توسعه انسانی سال آمارها برای کشور گزارش شده است برای مقایسه روشهای ممیزی خطی LDA درجه دوم QDA و ممیزی مخلوط MDA کشور بطور تصادفی برای مجموعه راهنما انتخاب شدند و نرخ خطای رده‌بندی نادرست برای مجموعه راهنما

6) Human Development Index



شکل   متغ رهای مم زی   متغ رهای مم زی اول و دوم بهتر کشورها را از هم جدا کرده‌اند

محاسبه شد سپس رای ررسی اعتبار مدلهای ارا ه شده کشور د گر ه عنوان مجموعه زمون در ن ر گرفته شده‌اند و نرخ خ ای رده ندی نادرست رای مشاهدات مجموعه زمون ا استفاده از توا مم زی رورد شده توس مجموعه راهنما محاسبه شده است هما ن ور که از جدول مشاهده می‌شود MDA ا سه ز رکلاس در هرکلاس دارای نرخ خ ای کمتری نسبت ه روشهای مم زی خ ی و درجه دوم است نا را ن از مم زی م خته رای رده ندی کشورها استفاده شده است

شکل نمودار مر و متغ رهای مم زی رای مم زی خ ی و مم زی م خته را رای مجموعه راهنما نشان می‌دهد هما ن ور که مشاهده می‌شود در MDA چون ز رکلاسها جا گز ن کلاسها شده‌اند نا را ن تعداد متغ رهای مم زی شتر از حالت خ ی است اعداد درون دا ره‌ها مشخص کننده مرکز ز رکلاسها می اشند و متغ رهای مم زی اول و دوم هتر از متغ رهای مم زی دوم و سوم کلاسها ز رکلاسها را از هم جدا کرده‌اند

کشورها ا توسعه انسانی الا

رژانت ن نت گو و ار ودا لمان ات ر ش اروگو ه اسپان ا استرال ا استونی اسلواکی اسلوونی ا تال ا امارات متحده عر ی انگلستان ا الات متحده مر کا ا رلند ا سلند ار ادوس اهاماس حر ن رونئ ی لژ ک پرتغال تر ن داد و تو اگو جمهوری چک دانمارک ژاپن سنگاپور سوئد سو س س شل فرانسه رژ م اشغالگر قدس فنلاند قبرس ق ر کاستار کا کانادا کرواسی کره‌جنو ی کو ت لوکزامبورگ لهستان مار ت وس مالت مالزی مجارستان مکز ک نروژ ن وزلند هلند هنگ کنگ ونان

کشورها ا توسعه انسانی متوس

ذر ا جان فر قای جنو ی لبانی اردن ارمنستان از کستان اکرا ن اکوادور اکوتور ال الجزا ر السالوادور اندونزی ا ران رز ل لغارستان ل ز ول وی وتسوانا پاپو پاراگو ه پاناما پرو تاج کستان تا لند ترکمنستان ترک ه تونس جاما کا جزا ر سل مان دم ن کا جمهوری دم ن کن چ ن روس ه روس ه سف د رومانی ز مباوه سامو ساوتومو سری لانکا سنت اوس ا سنت و ن ت سواز لند سور نام سور ه عر ستان عمان غنا ف ل پ ن ف جی قرق زستان قزاقستان کاپورد کامبوج کلمب ا کنگو کن ا کو ا گرجستان گرانادا گواتمالا گوانا جد د گو ان مالد و مراکش مصر مغولستان مقدون ه ملداوی م انمار لاتو ا ل توانی لبنان لسوتو ل بی ل تونی نامب ا ن کاراگو ه ونزولا و تنام هند هندوراس

کشورها ا توسعه انسانی پا ن

فر قای مرکزی نگولا ات و پی ار تره اوگاندا نگلادش ن ن وتان ورک نافاسو روندی پاکستان تانزان ا توگو چاد جمهوری مردمی لئو ج بوتی رو ندا زامب ا سنگال سودان س را ل ون کامرون کنگو کوته‌دو ری کوموروس گا ن گامب ا گ نه گ نه سا و ماداگاسکار مالاوی مالی مور تانی موزامب ک نپال ن جر ن جر ه وانوتو ها تی من

نت جه‌گ ر

هما ورکه گفته شد دل ل غ ر نرمال ودن کلاسها مم زی متداول کلاس ک توان کافی در جداسازی کلاسها را ندارند ک روش پارامتری را در مم زی م توان ا فرض داشتن توز نرمال م خته در کلاسها دست ورد ا ا ن فرض دو تعم م مهم در مم زی خ ی دست می د

کرانهای تصم م ن کلاسها غ ر خ ی و پ چ ده است

هنگام که مم زی ن دو کلاس اشد کاهش ا عاد ر خلاف LDA امکان‌پذ ر است

روش مم زی م خته اگر چه منجر ه محاسبات پ چ ده و استفاده از روشهای عددی جای تحل لی می‌گردد ولی ا توجه ه دسترسی ه را انه‌ها و نرم‌افزارهای دق ق می‌توان از روش MDA استفاده کرد

مراج

[1] Anderson, T. W.(1984), An Introduction to Multivariate Statistical Analysis, John Wiely, New York.

[2] Dempster, A. P.Laird, N. M. and Rubin, D. B.(1977), Maximum likelihood from incomplete data via EM algorithm(with discussion), R. Statis. Sac. B, 39:1-38.

[3] Hastie, T. Tibshirani, R. and Buja, A.(1994), Flexible Discriminant Analysis by Optimal Scoring, J. Amer. Statis. Assoc. 89, 1225-1270.

[4] Hastie, T. Tibshirani, R. (1996), Discriminant Analysis by Gaussian Mixtures, J. R. Statis. Soc. B, 58:155-176.

[5] Mardia, K. Kent, J. and Bibby, J.(1979), Multivariate Analysis, Academic Press, London.

[6] McLachlan, G. J. (1992), Discriminant Analysis and Statistica Pattern Recognition, John Wiley, New York.

[7] Taxt, T. Hjot, N. and Eikvil, L. (1991), Statistical classification using a linear mixture of multinormal probability densities, Pattern Recognition Letters, 12:731-737.

[8] UNDP. (2002), Human Development Report 2002, New York.

[9] Webb, A. (1999), Statistical Pattern Recognition, Arnold, London.