آنالیز ممیزی آمیخته درپیش بینی ورشکستگی

قاسم رکابدار[1]*، دکتر رحیم چینی پرداز[2]، بهاره سلیمانی[3]
[1و3] دانشگاه آزاد اسلامی واحد آبادان-گروه ریاضی
[2]دانشگاه شهید چمران اهواز -گروه آمار

چکیده

آنالیز ممیزی خطی بر مبنای نرمال بودن کلاس ها می باشد. درحالتی که کلاس ها نرمال نباشند برای رده بندی مؤثر بین کلاس ها و تعمیم ممیزی خطی می توان فرض کرد که کلاس ها از زیرکلاس هایی تشکیل شده اند که دارای توزیع نرمال هستند ممیزی در این حالت آمیخته می باشد و برآورد پارامترهای مدل آمیخته با الگوریتم EM امکان پذیر است در این حالت کران تصمیم بین کلاس ها غیرخطی است. برای رده بندی بین شرکت های ورشکسته و فعال در بورس تهران ممیزی خطی با ممیزی آمیخته مقایسه شد نتایج نشان داد که دقت رده بندی با ممیزی آمیخته بیشتر از ممیزی خطی است.

کلمات کلیدی: رده بندی، خوشه بندی، الگوریتم EM، آنالیز ممیزی خطی.

۱ مقدمه

رده بندی دو یا چند کلاس یا ممیزی از موضوعات مهم در آمار چند متغیره است که دارای کاربرد بسیاری در حوزه های گوناگون است. در حالت کلی ممیزی اختصاص یک یا چند مشاهده $\mathbf{x}$ با کلاس نامعلوم به جوامعی معلوم است. برای این کار فرض می شود مجموعه راهنما $\{\mathbf{x}_i, j\}_{i=1}^{n}$ که در آن $\mathbf{x}_i$ بردار d متغیره در فضای اقلیدسی $\Re^d$ بوده و $j \in \{1,2,...,J\}$ کلاس $\mathbf{x}_i$ و J نشان دهنده تعداد کلاس ها است. با استفاده از مجموعه راهنما پارامترها مدل برآورد و سپس با استفاده از پارامترهای برآورد شده، هدف بدست آوردن یک قاعده برای ممیزی به صورت یک تابع جهت پیش بینی مشاهده، جدید $\mathbf{x}$ به یکی از کلاس‌ها است.

روش سنتی و کلاسیک آماری برای رده بندی آنالیز ممیزی خطی[1] (LDA) است، این روش تا حدود زیادی توسعه یافته و معمولاً در متون آماری موجود است[۳] و [۶]. هر گاه جوامع نرمال باشند تابع ممیزی خطی نیز نرمال و بنابراین خطای ممیزی نیز براحتی قابل محاسبه خواهد بود. فرض نرمال بودن کلاس ها بندرت پیش می آید، در چنین صورتی تابع ممیزی داده ها را به صورت مؤثری از هم جدا نمی کند و خطای ممیزی افزایش

[1].Linear Discriminant Analysis

* عهده دار مکاتبات

خواهد یافت. یک طریق برای حل این مسأله فرض تعمیم LDA است به اینکه فرض شود کلاس‌ها شامل زیر کلاس‌هایی نرمال اند، یعنی کلاس‌ها دارای توزیع آمیخته اند و بنابراین ممیزی بکار رفته نیز آمیخته خواهد بود[۵] و [۸] و[۹].

۲ مدل آمیخته متناهی

چگالی آمیخته متناهی به صورت زیر تعریف می شود:

$$p(x)=\sum_{j=1}^{g}\pi_j p(\mathbf{x},\boldsymbol{\theta}_j) \qquad (۱)$$

که در این چگالی g تعداد مؤلفه های آمیخته و π_j نسبت های آمیختگی $(\sum_{j=1}^{g}\pi_j=1)$ است و $p(\mathbf{x},\boldsymbol{\theta}_j)$ توابع چگالی مؤلفه ها می باشند‌که وابسته به بردار پارامتری $\boldsymbol{\theta}_j$ هستند. در این صورت اگر n مشاهده به صورت $(\mathbf{x}_1,\dots,\mathbf{x}_n)$ داشته باشیم تابع درستنمائی به صورت زیر خواهد بود:

$$L(\Psi)=\prod_{i=1}^{n}\sum_{j=1}^{g}\pi_j p(\mathbf{x}_i\,|\,\boldsymbol{\theta}_j) \qquad (۲)$$

که Ψ مجمعه پارامترهای مدل $(\pi_1,\dots,\pi_g:\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_g)$ می باشد. بدیهی است ماکزیمم کردن رابطه (۲) از طریق مشتق گیری امکان پذیــر نیســت و باید با استفاده از روش های عددی ماکزیمم شود[۱۰].

۳ الگوریتم EM[1]

این الگوریتم دو مرحله ای برای برآورد پارامترها در هنگام ناکامل بودن مشاهدات کاربرد دارد[۴]. الگوریتم EM روشی تکراری تا رسیدن به همگرایی است که در هر تکرار، مرحله E (امید ریاضی) و M (ماکسیمم کردن) وجود دارد، در این مرحله با استفاده از این روش برآورد های ماکسیمم درستنمایی بدست می آیند.

۳-۱ برآورد ماکسیمم درستنمائی با EM

فرض شود $\{\mathbf{y}\}$ عبارت از کل داده های مشاهده شده و گم شده باشد یعنی بردار $\mathbf{x}_i$ با مقادیر گمشده اش افزوده شده باشد، بنابراین $\mathbf{y}'_i=(\mathbf{x}'_i,\mathbf{z}'_i)$ کامل است. در حالتی که با مدل های آمیخته روبـرو هستیم $\mathbf{z}_i$ یک بردار نشانگـــر برای نشان دادن کلاس مشاهــده می باشد یعنی $\mathbf{z}_i=(\mathbf{z}_{i1},\dots,\mathbf{z}_{gi})'$ است، بنابراین $\mathbf{z}_{ji}=1$ اگر $\mathbf{x}_i$ از مؤلفه j ام باشد و در جاهای دیگر صفر است.

1.Expectation-Maximation

درحالت کلی فرض می شود که تابع درستنمایی مشاهدات کامل $\mathbf{y}$ به صورت $g(\mathbf{y}\,|\,\Psi)$ است کـه شکل تابعی آن معلـــوم است. بنابراین تابع درستنمایی $p(\mathbf{x}\,|\,\Psi)$ را در حالت پیوسته می توان از $g(\mathbf{y}\,|\,\Psi)$ بوسیله انتگرال گیری روی همه مقادیر ممکن $\mathbf{y}$ از مجموعهٔ افزوده شده $\mathbf{x}$ به دست آورد:

$$L(\Psi)=p(\mathbf{x}\,|\,\Psi)=\int\prod_{i=1}^{n}g(\mathbf{x}_i,\mathbf{z}\,|\,\Psi)d\mathbf{z} \qquad (۳)$$

نتایجی که با روش EM برای دنباله ای از برآوردهای Ψ به دست آورده می شود، به صورت دنباله ای به شکل $\Psi^{(K)}$ از برآورد ابتدایی $\Psi^{(\circ)}$ است و شامل دو مرحله است[۴]:

۱ ـ **مرحله امید ریاضی** (E)**:** مرحله E از الگوریتم EM شامل امید ریاضی لگاریتم درستنمایی داده های کامل است، به شرط اینکه مقدار مشاهده شده $\mathbf{x}$ و مقدار صحیح پارامترهای در تکرار Kام الگوریتم $\Psi^{(K)}$ باشد، بنابراین تعریف می شود:

$$\begin{aligned}Q(\Psi;\Psi^{(K)})&\stackrel{\Delta}{=}E\left[\log g(\mathbf{y}\,|\,\Psi)\,|\,\mathbf{x},\Psi^{(K)}\right]\\&=\int\sum_{i=1}^{n}\log g(\mathbf{x}_i;\mathbf{z}_i\,|\,\Psi)p(\mathbf{z}\,|\,\mathbf{x}_i;\Psi^{(K)})\prod_{i=1}^{n}dz_i\end{aligned}$$

۲ ـ **مرحله ماکسیمم** (M)**:** مقدار جدیدی برای پارامتر $\Psi=\Psi^{(K+1)}$ که $Q(\Psi;\Psi^{(K)})$ را ماکسیمم کند محاسبه می شود. مراحل E و M تکرار می شوند تا همگرایی در تابع درستنمایی (۲) روی دهد. مراحل تا زمانی ادامه می یابد که شرط همگرائی برقرار باشد:

$$|\,L(\Psi^{(K+1)})-L(\Psi^{(K)})\,|<\varepsilon \qquad \forall\varepsilon>\circ$$

هرگاه تفاوت بیـن دو مقـدار ناچیز گردد الگوریتم متوقف می شود. تابع درستنمایی $L(\Psi)$ یک تابع اکیداً صعودی از Ψ می باشد زیرا:

$$\log(L(\Psi))=Q(\Psi,\Psi^{(K)})-H(\Psi,\Psi^{(K)})$$

که

$$\begin{aligned}H(\Psi,\Psi^{(K)})&=E[\log\{g(\mathbf{y}\,|\,\Psi)/p(\mathbf{x}\,|\,\Psi)\}\,|\,\mathbf{x},\Psi^{(K)}]\\&=\int\sum_{i}\log(p(\mathbf{z}_i\,|\,\mathbf{x},\Psi))p(\mathbf{z}_i\,|\,\mathbf{x},\Psi^{(K)})\prod_{i=1}^{n}dz_i\end{aligned}$$

چون $\log(x)\le x-1$ بنابراین خواهیم داشت:

$$H(\Psi^{(K+1)},\Psi^{(K)})-H(\Psi^{(K)},\Psi^{(K)})\le\circ$$

در اینجا $\Psi^{(K+1)}$ بگونه ای است که $Q(\Psi,\Psi^{(K)})$ را ماکسیمم می کند بنابراین در هر مرحله از الگوریتم EM خواهیم داشت:

$$\log(L(\Psi^{(K+1)}) \geq \log(L(\Psi^{(K)}) \qquad (۴)$$

۳-۲ برآورد پارامترهای چگالی آمیخته

برای داده هایی که کلاس آنها مشخص است بردار $\mathbf{y}$ مشاهده ای است که توسط کد کلاس افزوده شده است، بنابراین $\mathbf{y}' = (\mathbf{x}', \mathbf{z}')$ است که $\mathbf{z}$ یک بردار نشانگر به طول gاست که در موقعیت Kام یک است اگر $\mathbf{x}$ درمؤلفه Kام باشد در غیر اینصورت صفر است. درستنمایی $\mathbf{y}$ به شکل زیر تعریف می شود:

$$g(\mathbf{y} \mid \Psi) = p(\mathbf{x} \mid \mathbf{z}, \Psi) p(\mathbf{z} \mid \Psi) = p(\mathbf{x} \mid \boldsymbol{\theta}_K) \pi_K$$

که می تواند به شکل زیر نوشته شود:

$$g(\mathbf{y} \mid \Psi) = \prod_{j=1}^{g} \left[p(\mathbf{x} \mid \boldsymbol{\theta}_j) \pi_j \right]^{Z_j}$$

که Z_j در حالت $j = K$ برابر یک است و ودر غیراین صورت صفر است. درستنمایی $\mathbf{x}$ به شکل زیر است:

$$p(\mathbf{x} \mid \Psi) = \sum_{z} g(\mathbf{y} \mid \Psi) = \sum_{j=1}^{g} \pi_j p(\mathbf{x} \mid \boldsymbol{\theta}_j) \qquad (۵)$$

مشاهده می شود که درستنمایی $\mathbf{x}$ یک توزیع آمیخته متناهی است بنابراین با استفاده از (۵) می توان داده های آمیخته را به شکل داده های ناکامل تفسیر کرد که مقادیر گمشده در آن کد کلاس ها می باشند. برای n مشاهده $\mathbf{y}_1$، ...، $\mathbf{y}_n$ داریم:

$$g(\mathbf{y}_1, \ldots, \mathbf{y}_n \mid \Psi) = \prod_{i=1}^{n} \prod_{j=1}^{g} \left[\pi_j p(\mathbf{x}_i \mid \boldsymbol{\theta}_j) \right]^{Z_{ji}}$$

اگر از رابطه بالا لگاریتم گرفته شود:

$$\log g = \sum_{i=1}^{n} \sum_{j=1}^{g} Z_{ji} \left[\log \pi_j + \log p(\mathbf{x}_i \mid \boldsymbol{\theta}_j) \right] = \sum_{i=1}^{n} \sum_{j=1}^{g} Z_{ji} \log \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{g} Z_{ji} \log p(\mathbf{x}_i \mid \boldsymbol{\theta}_j)$$
$$= \sum_{i=1}^{n} \mathbf{Z}_i' \mathbf{L} + \sum_{i=1}^{n} \mathbf{Z}_i' u_i(\boldsymbol{\theta})$$

بردار $g, \mathbf{L}$ مؤلفه به صورت $\log \pi_j$ دارد و u_i دارای g مؤلفه به صورت $\log p(\mathbf{x}_i \mid \boldsymbol{\theta}_j)$ است و $\mathbf{Z}_i$ دارای مؤلفه هایی به صورت $j = 1, \ldots, g, Z_{ji}$ است که Z_{ji} متغیرهای نشانگر هستند که $Z_{ji} = 1$ اگر $\mathbf{x}_i$ در کلاس jام باشد و در غیر اینصورت صفر است.

۱ ـ مرحله E: در تکرار Kام Q به شکل زیر نمایش داده می شود:

$$
\begin{aligned}
Q\left(\Psi;\Psi^{(K)}\right) &= E\left[\log g\left(\mathbf{y}_1,\ldots,\mathbf{y}_n \mid \Psi^{(K)},\mathbf{x}_1,\ldots,\mathbf{x}_n\right)\right] \\
&= E\left[\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{L}+\sum_{i=1}^{n}\mathbf{Z}_i'u_i(\boldsymbol{\theta})\right] \\
&= \sum_{i=1}^{n}E\left(\mathbf{Z}_i' \mid \mathbf{x}_i,\Psi^{(K)}\right)\mathbf{L}+\sum_{j=1}^{n}E(\mathbf{Z}_i' \mid \mathbf{x}_i,\Psi^{(K)})u_i(\boldsymbol{\theta}) \\
&= \sum_{i=1}^{n}\mathbf{w}_i{}'\mathbf{L}+\sum_{i=1}^{n}\mathbf{w}_i'u_i(\boldsymbol{\theta})
\end{aligned}
$$

که $\mathbf{w}_i = E\left[\mathbf{Z}_i{}' \mid \mathbf{x}_i,\Psi^{(K)}\right]$ است و jامین مؤلفه آن احتمال این است، که $\mathbf{x}_i$ متعلق به کلاس یا مؤلفه jام باشد که از برآورد فعلی $\Psi^{(K)}$ بدست آورده می شود:

$$
\hat{W}_{ij}=\frac{\pi_j^{(K)}p\left(\mathbf{x}_i \mid \boldsymbol{\theta}_j^{(K)}\right)}{\sum_{m=1}^{g}\pi_m^{(K)}p\left(\mathbf{x}_i \mid \boldsymbol{\theta}_m^{(K)}\right)} \qquad (۶)
$$

۲ ـ مرحله M: همانطور که گفته شد این مرحله شامل ماکزیمم کردن Q با ملاحظه به Ψ است. به نوبت پارامترهای π_j، $\boldsymbol{\theta}_j$ بررسی می شوند Q با توجه به π_j ماکزیمم می شود به شرط اینکه $\sum_{j=1}^{g}\pi_j=1$ باشد با استفاده از ضرایب لاگرانژ داریم:

$$
F=Q-\lambda\left(\sum_{j=1}^{g}\pi_j-1\right)
$$

که معادل است با:

$$
F=\sum_{i=1}^{n}\mathbf{W}_i{}'\mathbf{L}+\sum_{i=1}^{n}\mathbf{W}_i'u_i(\boldsymbol{\theta})-\lambda\left(\sum_{j=1}^{g}\pi_j-1\right)
$$

اگر از F نسبت به π_j مشتق گرفته و حاصل برابر صفر قرار داده شود خواهیم داشت:

$$
\sum_{i=1}^{n}W_{ij}\frac{1}{\pi_j}-\lambda=0
$$

که با توجه به شرط $\sum_{j=1}^{g}\pi_j=1$ خواهیم داشت:

$$
\lambda=\sum_{j=1}^{g}\sum_{i=1}^{n}W_{ij}=n
$$

در نتیجه π_j به صورت زیر برآورد خواهد شد:

$$\hat{\pi}_j = \frac{1}{n}\sum_{i=1}^{n}\hat{W}_{ij} \tag{۷}$$

بنابراین برآورد نسبت آمیختگی برای مؤلفهٔ j ام میانگین وزنهای آمیختگی برای مؤلفه j ام است[۱۰].

٤ آنالیز ممیزی آمیخته[1]

فرض کنید تابع چگالی در هر کلاس چگالی آمیخته نرمال باشد که در آن هر مؤلفه دارای میانگین مربوط به خود و ماتریس کواریانس مشترک با بقیه مؤلفه ها باشد، بنابراین در کلاس j ام آمیخته نرمال به صورت زیر است:

$$p_j(\mathbf{x}) = \sum_{r=1}^{R_j}\pi_{jr}\phi_{jr}(\boldsymbol{\mu}_{jr},\boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}}\sum_{r=1}^{R_j}\pi_{jr}\exp\left\{-\frac{1}{2}D^2(\boldsymbol{\mu}_{jr},\boldsymbol{\Sigma})\right\} \tag{۸}$$

که R_j تعداد زیرکلاس‌های گروه j ام و π_{jr} نسبت های آمیختگی هر یک از مؤلفه های نرمال است ($\sum_{r=1}^{R_j}\pi_{jr} = 1$). D^2 فاصله ماهالانوبیس بین بردار $\mathbf{x}$ و میانگین زیر کلاس r ام از کلاس j ام است که به صورت زیر تعریف می شود:

$$D^2(\boldsymbol{\mu}_{jr},\boldsymbol{\Sigma}) = (\mathbf{x}-\boldsymbol{\mu}_{jr})'\boldsymbol{\Sigma}(\mathbf{x}-\boldsymbol{\mu}_{jr})$$

در چنین حالتی احتمال پسین کلاس j ام در صورتیکه $\mathbf{x}$ مشاهده شده باشد به صورت زیر است:

$$P(G=j\mid\mathbf{X}=\mathbf{x}) = \frac{\sum_{r=1}^{R_j}\pi_{jr}\phi_{jr}(\boldsymbol{\mu}_{ir},\boldsymbol{\Sigma})}{\sum_{j=1}^{J}\sum_{r=1}^{R_j}\pi_{jr}\phi_{jr}(\boldsymbol{\mu}_{jr},\boldsymbol{\Sigma})} \tag{۹}$$

و لگاریتم درستنمایی شرطی به صورت زیرخواهد بود:

$$\ell_{mix}(\boldsymbol{\mu}_j,\boldsymbol{\Sigma},\pi_{jr}) \propto \sum_{j=1}^{J}\sum_{g_i=j}\log\Big(\sum_{r=1}^{R_j}\pi_{jr}\exp\left\{-\frac{1}{2}D^2(\mathbf{x}_i,\boldsymbol{\mu}_{jr})\right\}\Big) - \frac{n}{2}\log|\boldsymbol{\Sigma}| \tag{۱۰}$$

در اینجا مشاهده جدید $\mathbf{x}_0$ با ماکزیمم کردن رابطه (۱۰) به کلاس j ام نسبت داده می شود اگر :

$$P(G=j\mid\mathbf{X}=\mathbf{x}_0) = \max_{e} P(G=\ell\mid\mathbf{X}=\mathbf{x}_0) \tag{۱۱}$$

[1] Mixture Discriminant Analysis

درممیزی آمیخته پارامترهای مدل $(j=1,...,J, r=1,...,R_j)\Psi=\{R_j,\pi_{jr},\boldsymbol{\mu}_{jr},\boldsymbol{\Sigma}\}$ می باشند که باید از طریق مجموعه راهنما برآورد شوند. هاستی و تیبشیرانی از الگوریتم EM جهت برآورد پارامترهای مدل آمیخته در ممیزی استفاده کردند که به صورت زیر هر تکرار الگوریتم می باشد[۵]:

مرحله امید ریاضی E:

با فرض اینکه پارامترها درست داده شده اند با استفاده از رابطه(۶) وزن یا احتمال اینکه مشاهده $\mathbf{x}_i$ متعلق به زیر کلاس r از کلاس j ام باشد (C_{jr})برای همه مشاهدات مجموعه راهنما در کلاس j محاسبه می شود:

$$w_{ijr}=\frac{\pi_{jr}\phi(\mathbf{x}_i,\boldsymbol{\mu}_{jr},\boldsymbol{\Sigma})}{\sum_{\ell=1}^{R_j}\pi_{j\ell}\phi(\mathbf{x}_i,\boldsymbol{\mu}_{j\ell},\boldsymbol{\Sigma})} \tag{۱۲}$$

مرحله ماکسیمم کردن M:

با استفاده از وزن های برآورد شده در مرحله E برآوردهای ماکسیمم درستنمایی برای پارامترهای هر مؤلفه نرمال در هر کلاس محاسبه می شود در اینجا $\sum_{g_i=j}$ مجموع مشاهدات در کلاس j ام است :

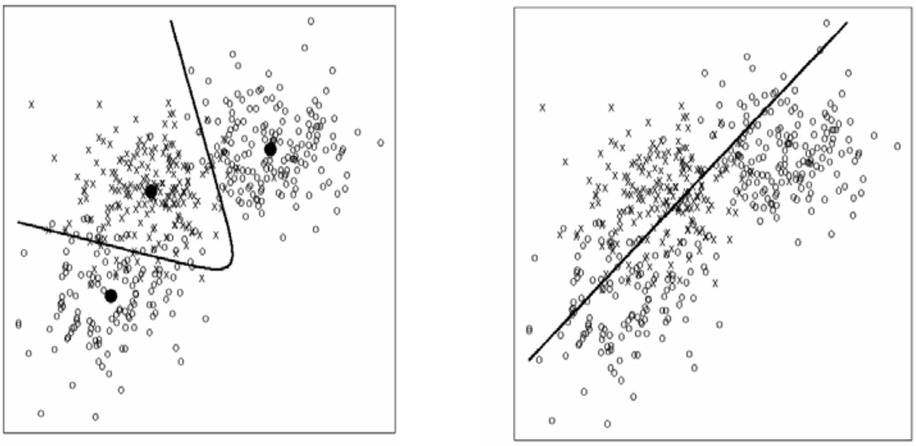
$$\hat{\pi}_{jr}\propto\sum_{g_i=j}w_{ijr},\qquad \sum_{r=1}^{R_j}\pi_{jr}=1 \tag{۱۳}$$

$$\hat{\boldsymbol{\mu}}_{jr}=\frac{\sum_{g_i=j}w_{ijr}\mathbf{x}_i}{\sum_{g_2=i}w_{ijr}} \tag{۱۴}$$

$$\hat{\boldsymbol{\Sigma}}=\frac{1}{n}\sum_{j=1}^{J}\sum_{g_i=j}\sum_{r=1}^{R_j}w_{ijr}\left(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_{jr}\right)\left(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_{jr}\right)' \tag{۱۵}$$

با توجه به این که وزن w_{ijr} خود تابعی از $\boldsymbol{\mu}_{jr}$ و $\boldsymbol{\Sigma}$ است رابطه های (۱۴) و (۱۵) در رابطهٔ (۱۲) جایگزین شده تا وزن های بعدی محاسبه شوند این مراحل آنقدر تکرار می شوند تا همگرایی در لگاریتم درستنمایی شرطی (۱۰) روی دهد یعنی اگر در مرحله به مرحله بعدی رابطه (۱۰)تغییر کمی داشته باشد الگوریتم متوقف می شود وبرآوردهای نهایی بدست آورده می شوند.

تکرارهای الگوریتم EM به اندازه خوشه ها R_j و مقادیر آغازین برای $\boldsymbol{\mu}_{jr}$ و $\boldsymbol{\Sigma}$ و وزن های w_{ijr} بستگی دارند. به این منظور می توان از الگوریتم خوشه بندی میانگین K ام یک مقدار ثابت برای خوشه ها تعیین کرد و سپس از این الگوریتم برای برآورد مراکز زیر کلاس‌های $\hat{\boldsymbol{\mu}}_{jr}$ در هر کلاس استفاده شود آنگاه برای همه مشاهدات در کلاس jام w_{ijr} برابر یک است اگر $\hat{\boldsymbol{\mu}}_{jr}$ نزدیکترین مرکز به $\mathbf{x}_i$ باشد و در غیر این صورت صفر است.



شکل ۱: کران تصمیم بین دو کلاس سمت چپ MDA سمت راست LDA

شکل (۱) یک مثال ساده شبیه سازی شده را نشان می دهد، در کلاس اول ۴۰ مشاهده دارای توزیع نرمال دو متغیره با میانگین (۰ و ۰) می باشند و ۴۰ مشاهده دیگر دارای توزیع نرمال دو متغیره با میانگین (۴ و۴) هستند و در کلاس دوم مشاهدات دارای توزیع نرمال با میانگین (۲ و۲) هستند و ماتریس کواریانس در هر دو کلاس به صورت یکانی $\Sigma = I_2$ است. همان طور که مشاهده می شود در ممیزی آمیخته کران تصمیم بین کلاس ها غیر خطی است و کلاس ها بهتر از ممیزی خطی از هم جدا شده اند این تعمیم مهمی است که با فرض نرمال آمیخته برای چگالی کلاس ها در ممیزی خطی روی می دهد.

۵ پیش بینی ورشکستگی

عدم اطمینان از بقاء شرکت های موجود در بورس، یکی از عواملی است که ممکن است سرمایه گذاران را از سرمایه گذاری در بورس منصرف کند زیرا ممکن است شرکت در همان سال ورشکست شده و از بورس خارج شود لذا باید با طراحی و پیشنهاد الگوهای مناسب آماری جهت پیش بینی شرکت هایی که در حال ورشکستگی هستند، می توان به سرمایه گذاران این اطمینان را داد تا با ریسک کمتری در بورس سرمایه گذاری کنند.

در ادبیات مالی اصطلاح شرکتهای ورشکسته را این‌طور تعریف می‌کنند: "واحدهای تجاری که عملیات خود را به علت واگذاری یا توقف انجام عملیا ت جاری با زیان توسط بستانکاران، متوقف نمایند"[۱]. پیش‌بینی ورشکستگی با استفاده از روش های مختلفی صورت می‌پذیرد، که از میان روش های مزبور، روش تجزیه و تحلیل نسبت ها و روش تجزیه و تحلیل ریسک بازار از اعتبار بیشتری برخوردار است. در روش تجزیه‌وتحلیل ریسک بازار، احتمال وقوع ورشکستگی شرکت ازطریق تغییراتی که در ریسک بازار (مثل واریانس نرخ بازده یک سهم و ریسک سیستماتیک) رخ می‌دهد، تخمین زده می‌شود. در روش تجزیه‌وتحلیل نسبت ها، احتمال

وقوع ورشکستگی به‌وسیله یک گروه از نسبت های مالی که توسط صاحب نظران با هم ترکیب شده‌اند تخمین‌زده می‌شود.

۵-۱ مدل آلتمن

آلتمن نخستین فردی می باشد که مدل های پیش بینی چند متغیره را مطرح ساخته است[۲]. او با به کار گیری روش آنالیز ممیزی خطی و استفاده از نسبت های مالی به عنوان متغیرهای مستقل بدنبال پیش بینی ورشکستگی در بنگاه ها بود. او با این روش از میان بیست و دو نسبت مالی برای پیش بینی ورشکستگی پنج نسبت را گزینش کرد که دارای بهترین عملکرد در میان دیگر نسبت های مالی هستند این نسبت ها عبارتند از:۱)سرمایه در گردش تقسیم بر کل دارایی ها. ۲)سود انباشته تقسیم بر کل دارایی ها. ۳)سود قبل از بهره و مالیات تقسیم بر کل دارایی ها. ۴)ارزش دفتری حقوق صاحبان سهم تقسیم بر ارزش دفتری کل بدهی ها. ۵) فروش تقسیم بر کل دارایی ها. در سال های بعد انتقاداتی به مدل آلتمن وارد شد از جمله اینکه این مدل برای شرکت های پذیرفته شده در بورس و تولیدی مناسب است و برای شرکت های خدماتی مفید نیست. به عبارت دیگر برای مدل آلتمن باید شرکت های همگن را استفاده نمود. با وجود گذشت نزدیک به چهار دهه از تحقیق آلتمن، همچنان نسبت های مالی مدل وی مورد علاقه بسیاری از پژوهشگران برای پیشبینی ورشکستگی بوده است.

۵-۲ پیش بینی ورشکستگی باMDA

آنالیز ممیزی خطی روش سنتی آماری جهت رده بندی می باشد. اما همانطور که گفته شد این روش برای موقعی که گروه ها دارای توزیع نرمال یا گروه ها همگن باشند مناسب است. در پیش بینی ورشکستگی نیز معمولاً گروه ها همگن و از یک صنعت خاص انتخاب می شوند. استفاده از آنالیز ممیزی آمیخته نیازی به همگن بودن گروه ها ندارد یعنی با تفکیک گروه ها به زیر گروه ها می توان بر این مشکل فائق آمد.

جامعه آماری این پژوهش کلیه شرکت های پذیرفته شده در بورس اوراق بهادار تهران می باشد که در بین سال های ۱۳۷۸ تا ۱۳۸۴ صورت های مالی خود را به بورس ارایه کرده بودند. برای گردآوری اطلاعات از بانک های اطلاعاتی مربوط به بورس اوراق بهادار و نرم افزار دنا سهم و اطلاعات موجود در اینترنت استفاده شده است.

روش	گروه واقعی	گروه پیش بینی شده		درصد دقت رده بندی	درصد دقت کلی
		فعال	ورشکسته		
LDA	فعال	۹۶	۴	۹۶	۹۰/۳
	ورشکسته	۱۰	۳۴	۷۷/۳	
MDA(R=2)	فعال	۹۸	۲	۹۸	۸۹/۶
	ورشکسته	۱۳	۳۱	۷۱	
MDA(R=3)	فعال	۹۸	۲	۹۸	۹۲/۳۴
	ورشکسته	۹	۳۵	۸۰	
MDA(R=4)	فعال	۹۷	۳	۹۷	۹۱/۷
	ورشکسته	۹	۳۵	۸۰	

جدول ۱: درصد دقت رده بندی در بورس تهران با ممیزی خطی (LDA) و ممیزی آمیخته(MDA)

به دلیل این که منبع مناسبی جهت مشخص کردن شرکت های ورشکسته در بورس تهران یافت نشد در این تحقیق شرکت هائی به عنوان نمونه شرکت های دارای وضعیت بحرانی در نظر گرفته شدند که بین سال های ۱۳۷۸ تا ۱۳۸۴ نماد آنها حداقل دو سال بسته بوده است. و همچنین در بین این سال ها نماد آنها همچنان بسته مانده است با این روش ۴۴ شرکت به عنوان نمونه شرکت های غیرفعال در نظر گرفته شد. برای این شرکت ها نسبت مالی سال قبل از بحران مالی به عنوان داده ها تحلیل نهایی در نظر گرفته شد. بدلیل عدم محدودیت در انتخاب شرکت های فعال در بورس بطور تصادفی از شرکت های همه صنایع که به صورت مستمر سهام ماهانه آنها معامله شده انتخاب گردید که در نهایت ۱۰۰ شرکت برای شرکت های غیرورشکسته یا فعال بورس انتخاب شد[۱].

برای پیش بینی ورشکستگی با مدل های ممیزی خطی(LDA) و ممیزی آمیخته(MDA) از نرم افزار SPLUS/2000 استفاده شده است. نسبت های مالی آلتمن برای ۱۰۰ شرکت فعال و ۴۴ شرکت ورشکسته به عنوان متغیر مستقل محاسبه شدند و وضعیت شرکت به عنوان متغیر وابسته(ورشکسته=۱ و فعال=۲) در نظر گرفته شد. جدول(۱) نشان دهنده وضعیت رده بندی این شرکت ها با دو مدل است برای انتخاب تعداد زیر کلاس ها در MDA درصد دقت به عنوان ملاک در نظر گرفته شد. همانطور که از جدول(۱) دیده می شود دقت LDA برای رده بندی شرکت های فعال ۹۶ درصد و برای شرکت های ورشکسته ۷۷/۳ درصد است در حالت کلی دقت رده بندی LDA برای پیش بینی ورشکستگی ۹۰/۳ درصد است. از میان مدل های آمیخته مدل MDA با سه زیرکلاس در هر کلاس دارای بهترین درصد دقت می باشد برای این مدل ۹۸ درصد شرکت های فعال و ۸۰ درصد شرکت های ورشکسته درست رده بندی شده اند و درصد دقت کلی ۹۲/۳۴ درصد است. بنابراین برای پیش بینی ورشکستگی در بورس تهران مدل ممیزی آمیخته دقت بیشتری نسبت به ممیزی خطی دارد.

٦ نتیجه گیری

هنگامی که کلاس ها با چگالی نرمال آمیخته مدل بندی می شوند کران تصمیم بین کلاس ها غیر خطی خواهد بود و در حالت نرمال نبودن کلاس ها این روش منجر به رده بندی بهینه نسبت به ممیزی خطی خواهد شد. روش ممیزی آمیخته اگر چه منجر به محاسبات پیچیده و استفاده از روش های عددی بجای تحلیلی می گردد ولی با توجه به دسترسی به رایانه ها با سرعت بالا و نرم افزارهای قوی می توان از روش MDA استفاده کرد. نتایج به دست آمده برای پیش بینی ورشکستگی در بورس تهران بیانگر کارآئی بهتر MDA بر LDA می باشد.

منابع

[۱] رکابدار، قاسم(۱۳۸۶). مدل آمیخته برای پیش بینی ورشکستگی. معاونت پژوهشی دانشگاه آزاد اسلامی آبادان.

[2] Altman, E.I.(1968) Financial Ratios, Discriminant Analysis and the Prediction of Corporate bankruptc. Journal of Finance, September,.589-609.

[3] Anderson, T. W, (1984). An Introduction to Multivariate Statistical Analysis. Johnwiley , New york.

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

[5] Hastie , T. And Tibshirani, R (1996).Discriminant analysis by gaussian mixtures.J. R. Statist. Soc. B, 58: 155 - 176.

[6] Mardia , k. , Kent, Jand Bibby , J. (1979).Multivariate Analysis. Academic press.

[7] Mclachlan , G. J. (1992).Discriminant Analysis and Statistical Pattern Recognition.John wiley , New york.

[8] McLachlan, G.J. (1999). Finite Mixture Models. New York:Wiley.

[9] Taxt, T. Hot, N. and Eikvil, L. (1991).Statistical Classification using a linear Mixtures of Multinormal Probability densities.Pattern Recognition Letters , 12; 731 - 737.

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