مدل آمیختهٔ بیزی برای رده‌بندی داده‌های دقیق و نادقیق

ایمانه خدایاری صمغ‌آبادی[۱]، فرزاد اسکندری[۲]

تاریخ دریافت: ۱۳۹۳/۱۲/۱۷
تاریخ پذیرش: ۱۳۹۵/۷/۱۷

چکیده:

رده‌بندی داده‌های دقیق تا کنون با روش‌های مختلف و در ابعاد وسیعی مورد بررسی و تحلیل قرار گرفته است، اما داده‌هایی که برای رده‌بندی مورد استفاده قرار می‌گیرند همیشه مقدار مشخص و دقیقی ندارند. از آن‌جا که نوع مقیاس داده‌ها متفاوت است، مقدار داده ممکن است در یک بازه قرار گیرد که در این صورت، مسئلهٔ رده‌بندی داده‌های نادقیق مطرح می‌شود. در سال‌های اخیر با فرض نرمال بودن توزیع حاکم بر داده‌های نادقیق، برآوردهای مختلفی برای میانگین و واریانس این توزیع ارائه‌شده است. در این مقاله با فرض این‌که توزیع حاکم بر داده‌های نادقیق توزیع نرمال دومتغیره باشد، با روش ماکسیمم درست‌نمایی بر روی مقادیر دو سر بازه داده‌های نادقیق، میانگین و واریانس این توزیع را برآورد کرده‌ایم. سپس با استفاده از رده‌بندی سادهٔ بیزی، یک مدل آمیختهٔ بیزی برای رده‌بندی داده‌های دقیق و نادقیق ارائه کرده‌ایم. همچنین دقت و کارایی مدل ارائه‌شده بررسی شده است.

واژه‌های کلیدی: رده‌بندی داده‌ها، رده‌بندی سادهٔ بیزی، صفت عددی نادقیق، برآورد ماکسیمم درست‌نمایی، دقت.

۱ مقدمه

یکی از کاربردهای یادگیری ماشین،[۳] رده‌بندی[۴] است. در هر روش رده‌بندی، از یک الگوریتم یادگیری استفاده می‌شود تا به‌کمک مجموعه‌ای از نمونه‌های آموزشی، مدلی را برای رده‌بندی داده‌ها به دست آورد. سپس می‌توان از مدل به دست آمده در مراحل بعدی برای پیش‌بینی ردهٔ نمونه‌هایی استفاده کرد که در نمونه‌گیری‌های مختلف ایجاد می‌شود. بنا بر این یادگیری آن یک یادگیری باناظر[۵] است. در یادگیری باناظر، مجموعه‌ای از نمونه‌های آموزشی وارد مدل رده‌بندی می‌شوند تا تابعی مناسب برای رده‌بندی نمونه‌های جدید ارائه کند. رده‌بندی درخت تصمیم، شبکهٔ عصبی، نزدیک‌ترین همسایه، ماشین بردار پشتیبان و رده‌بندی سادهٔ بیزی از جمله روش‌های رده‌بندی داده‌ها می‌باشند؛ اما داده‌هایی که برای رده‌بندی مورد استفاده قرار می‌گیرند، همیشه مقدار مشخص و دقیقی ندارند. در واقع نوع مقیاس داده‌ها ممکن است متفاوت بوده؛ در بازه‌ای از مقادیر وجود داشته باشد. به‌عنوان مثال، برای سؤال‌هایی از قبیل «در هفته چند ساعت تلویزیون تماشا می‌کنید؟» پاسخ دقیقی نداریم؛ ولی اگر جواب را به‌صورت بازه در نظر بگیریم، پاسخ‌گویی به این سؤال آسان‌تر می‌شود. بنا بر این علاوه بر داده‌هایی که مقدار مشخص و دقیقی دارند، داده‌هایی نیز هستند که متعلق به بازه‌ای از مقادیرند و ما آن‌ها را داده‌های نادقیق می‌نامیم.

برای اولین بار بیلارد و دیدی [۵] یک مدل رگرسیونی برای رده‌بندی داده‌های نادقیق با استفاده از نقاط مرکزی داده‌های نادقیق ارائه کردند، اما استفاده از این مدل باعث از بین رفتن داده‌های اصلی می‌شود. برای رفع این مشکل، کاروالیو و همکاران [۶] مدل آن‌ها را با به‌کارگیری دامنهٔ داده‌های نادقیق توسعه دادند. علاوه بر مدل‌های رگرسیونی، مدل‌های خوشه‌بندی نیز مورد توجه قرار گرفتند [۷، ۸، ۹، ۱۰]. همچنین کین و همکاران [۱۱] رده‌بندی داده‌های نادقیق را با درخت تصمیم انجام دادند؛ اما مدل آن‌ها فقط یک نوع داده‌های نادقیق را رده‌بندی می‌کند. بنا بر این کین و همکاران [۱۳] با در نظر

[۱] کارشناسی ارشد علوم کامپیوتر موسسهٔ آموزش عالی غیاث‌الدین جمشید کاشانی

[۲] دانشیار گروه آمار دانشگاه علامه طباطبایی

[۳] machine learning

[۴] classification

[۵] supervised learning

گرفتن توزیع نرمال یک‌متغیره برای داده‌های نادقیق، برآوردی برای میانگین و واریانس آن به دست آوردند و با استفاده از رده‌بندی سادهٔ بیزی، رده‌بندی بیزی جدیدی را برای انواع داده‌های دقیق و نادقیق ارائه کردند. لورادهماخر و بیلارد [۱۴] برآوردهایی را برای انواع داده‌های نمادین[۶] با روش ماکسیمم درست‌نمایی ارائه کردند. داده‌های نمادین، توصیف افراد و گروه‌ها را ممکن می‌سازد و شامل داده‌هایی با مقادیر وزن‌دار یا بدون وزن، بازه‌ای، بافت‌نگاشتی، چندمقداری و ... است. آن‌ها با توجه به این‌که داده‌های نمادین یک ساختار داخلی دارند که در داده‌های کلاسیک وجود ندارد، پارامترهای میانگین و واریانس را برای داده‌های بافت‌نگاشت، بازه‌ای و ... با روش ماکسیمم درست‌نمایی برآورد کرده‌اند.

در این مقاله توزیع داده‌های نادقیق را توزیع نرمال دومتغیره فرض کرده‌ایم و برآورد میانگین و واریانس مقادیر کران پایین و بالای آن را با روش ماکسیمم درست‌نمایی به دست آورده‌ایم. سپس با استفاده از رده‌بندی سادهٔ بیزی، یک مدل آمیختهٔ بیزی برای رده‌بندی داده‌های دقیق و نادقیق با همان دقت مدل کین و همکاران [۱۳] ارائه می‌کنیم. این مقاله ۶ بخش دارد. در بخش اول، مقدمه‌ای از رده‌بندی داده‌های نادقیق بیان شده است. در بخش دوم، روش رده‌بندی سادهٔ بیزی را معرفی می‌کنیم. در بخش سوم و چهارم با روش ماکسیمم درست‌نمایی، برآورد میانگین و واریانس یک صفت عددی دقیق و نادقیق را به دست می‌آوریم. در بخش پنجم، الگوریتم آمیختهٔ بیزی را برای رده‌بندی داده‌های عددی دقیق و نادقیق ارائه می‌کنیم. همچنین نرمال بودن داده‌های آموزشی بررسی می‌شود. در بخش ششم نتایج عددی مدل آمیختهٔ بیزی را به‌صورت شبیه‌سازی و در ساختار یک مثال عددی مورد بررسی و دقت مدل پیشنهادی نیز مورد ارزیابی قرار خواهد گرفت.

۲ رده‌بندی سادهٔ بیزی

یکی از روش‌های کاربردی یادگیری ماشین، روش رده‌بندی سادهٔ بیزی است. این روش برای رده‌بندی نمونهٔ جدید، با داشتن مقادیر صفت‌های $A_1, ..., A_n$ که توصیف‌کنندهٔ نمونهٔ جدید است، محتمل‌ترین رده (C_k) را به‌صورت زیر شناسایی می‌کند [۱۶]:

$$P(C_k \mid A_1, ..., A_n) = \frac{P(C_k)P(A_1, ..., A_n|C_k)}{P(A_1, ..., A_n)} \tag{۱}$$

رده‌بندی سادهٔ بیزی، احتمال شرطی رده‌ها را با فرض این‌که صفت‌ها از یکدیگر مستقل شرطی‌اند، به دست می‌آورد. بنا بر این رابطهٔ (۱) می‌تواند به‌صورت زیر جایگزین شود:

$$P(C_k \mid A_1, ..., A_n) = \frac{P(C_k)\prod_{i=1}^{n} P(A_i|C_k)}{P(A_1, ..., A_n)} \tag{۲}$$

از آن‌جاکه $P(A_1, ..., A_n)$ برای هر C_k ثابت است، برای به دست آوردن محتمل‌ترین رده، کافی است رده‌ای را انتخاب کنیم که رابطهٔ (۲) را ماکسیمم کرده است.

فرض کنید نمونهٔ T_j، دارای n صفت $A_1, ..., A_n$ باشد که از بین آن p صفت، صفت عددی دقیق و $n-p$ صفت باقی‌مانده، صفت عددی نادقیق‌اند. احتمال این‌که نمونهٔ T_j در ردهٔ C_k باشد، به‌صورت زیر به دست می‌آید:

$$\begin{aligned} P(C_k|T_j) =& arg \max_{C_k \in C} P(C_k) \\ & \times \prod_{i=1}^{p} P(A_i|C_k) \times \prod_{i=p+1}^{n} P(A_i|C_k) \end{aligned} \tag{۳}$$

برای همهٔ رده‌ها $P(C_k|T_j)$ محاسبه و T_j بر اساس رده‌ای با بیشترین احتمال، رده‌بندی می‌شود.

در این مقاله ردهٔ C_k به‌ازای $k = 1, ..., p$ معین فرض شده است.

تعریف ۱.۲. A_i صفت عددی دقیق[۷] است هرگاه هر یک از مقادیر عددی A_{ij} دقیق باشد. A_{ij}، jامین مقدار از صفت A_i به‌ازای $j = 1, ..., m$ است.

تعریف ۲.۲. A_i صفت عددی نادقیق[۸] است هرگاه هر یک از مقادیر عددی A_{ij} نامشخص و متعلق به بازه‌ای از مقادیر $[a_{ij}, b_{ij}]$ به‌ازای $j = 1, ..., m$ باشد.

تعریف ۳.۲. نمونه، مقادیر صفت‌های $A_1, ..., A_n$ که در کنار یکدیگرند و اطلاعاتی در بارهٔ قرار گرفتن در یک رده را نشان می‌دهند.

[۶] symbolic data

[۷] certain numerical attribute

[۸] uncertain numerical attribute

۳ برآورد ماکسیمم درست‌نمایی صفت عددی دقیق

برآورد ماکسیمم درست‌نمایی روشی برای برآورد کردن پارامترهای یک مدل آماری است. فرض کنید داده‌ها از توزیع نرمال پیروی می‌کنند و میانگین و واریانس آن مجهول است. با استفاده از ماکسیمم درست‌نمایی و با داشتن اطلاع مربوط به نمونه‌ای محدود از جمعیت، می‌توان میانگین و واریانس توزیع آن را به دست آورد. این روش، میانگین و واریانس را مجهول در نظر می‌گیرد و مقادیری را به آن‌ها نسبت می‌دهد که با توجه به اطلاع موجود، محتمل‌ترین حالت باشد.

اگر A_i صفت عددی دقیق باشد و از توزیع نرمال تبعیت کند، برای رده‌بندی سادهٔ بیزی می‌توان از توزیع نرمال زیر استفاده کرد [۱۵]:

$$P(A_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \exp\left\{-\frac{(x-\mu_{ik})^2}{2\sigma_{ik}^2}\right\},$$

که x نشان‌دهندهٔ مقدار صفت A_i از نمونهٔ جدید است. برآورد میانگین و واریانس این توزیع به‌کمک:

$$\begin{aligned}\hat{\mu}_{ik} &= \frac{1}{m}\sum_{j=1}^{m} a_{ij}\\ \hat{\sigma}_{ik}^2 &= \frac{1}{m-1}\sum_{j=1}^{m}(a_{ij}-\mu_{ik})^2\end{aligned}$$

که در آن m تعداد نمونه‌های ردهٔ C_k و a_{ij} مقادیر صفت عددی دقیق A_i به‌ازای $j=1,\dots,m$ است.

۴ برآورد ماکسیمم درست‌نمایی صفت عددی نادقیق

اگر A_i صفت عددی نادقیق باشد و مقادیر کران پایین و بالای آن از توزیع نرمال تبعیت کند، با تعریف میانگین $E[a_{ij}] = \mu_{i1}$ و واریانس $var(a_{ij}) = \sigma_{i1}^2$ برای مقادیر کران پایین $(a_{i1}, a_{i2}, ..., a_{im})$ و میانگین $E[b_{ij}] = \mu_{i2}$ و واریانس $var(b_{ij}) = \sigma_{i2}^2$ برای مقادیر کران بالا $(b_{i1}, b_{i2}, ..., b_{im})$، می‌توانیم توزیع نرمال دومتغیره برای صفت عددی نادقیق A_i را به‌صورت زیر در نظر بگیریم:

$$A_i := \begin{pmatrix} a_{ij} \\ b_{ij} \end{pmatrix} \sim N(\mu, \Sigma), \quad j = 1, ..., m$$

در این توزیع، کوواریانس، بردار میانگین μ و ماتریس واریانس-کوواریانس Σ به‌صورت زیر است:

$$Cov[a_{ij}, b_{ij}] = \rho_{i12}\sigma_{i1}\sigma_{i2}, \quad \mu = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} \sigma_{i1}^2 & \sigma_{i1}\sigma_{i2}\rho_{i12} \\ \sigma_{i1}\sigma_{i2}\rho_{i12} & \sigma_{i2}^2 \end{pmatrix}.$$

بنا بر این احتمال شرطی صفت عددی نادقیق A_i در رده C_k به‌صورت زیر به دست می‌آید:

$$\begin{aligned}P(A_{ik}|C_k) =& \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}}\\ &\times \exp\left(-\frac{1}{2}(A_{ik}-\mu)'\Sigma^{-1}(A_{ik}-\mu)\right)\end{aligned} \tag{4}$$

که دترمینان ماتریس واریانس-کوواریانس عبارت است از:

$$|\Sigma| = \sigma_{i1}^2\sigma_{i2}^2(1-\rho_{i12}^2).$$

با جایگذاری بردار میانگین، دترمینان و وارون ماتریس واریانس-کوواریانس در رابطهٔ (۴) داریم:

$$\begin{aligned}P(A_i|C_k) =& \frac{1}{2\pi\sigma_{i1}\sigma_{i2}\sqrt{1-\rho_{i12}^2}} \exp\Bigg\{ - \frac{1}{2(1-\rho_{i12}^2)}\\ &\times \left[\left(\frac{a_{ij}-\mu_{i1}}{\sigma_{i1}}\right)^2 + \left(\frac{b_{ij}-\mu_{i2}}{\sigma_{i2}}\right)^2\right.\\ &\left. - 2\rho_{i12}\left(\frac{a_{ij}-\mu_{i1}}{\sigma_{i1}}\right)\left(\frac{b_{ij}-\mu_{i2}}{\sigma_{i2}}\right)\right]\Bigg\}\end{aligned} \tag{5}$$

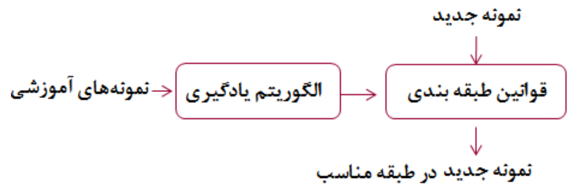
از رابطهٔ (۵) تابع درست‌نمایی زیر نتیجه می‌شود:

$$\begin{aligned}L(\mu, \Sigma; A_i) =& \prod_{j=1}^{m}\Bigg\{\left(\frac{1}{2\pi\sigma_{i1}\sigma_{i2}\sqrt{1-\rho_{i12}^2}}\right) exp\Bigg\{ - \frac{1}{2(1-\rho_{i12}^2)}\\ &\times \left[\left(\frac{a_{ij}-\mu_{i1}}{\sigma_{i1}}\right)^2 + \left(\frac{b_{ij}-\mu_{i2}}{\sigma_{i2}}\right)^2\right.\\ &\left. - 2\rho_{i12}\left(\frac{a_{ij}-\mu_{i1}}{\sigma_{i1}}\right)\left(\frac{b_{ij}-\mu_{i2}}{\sigma_{i2}}\right)\right]\Bigg\}\Bigg\}\end{aligned} \tag{6}$$

برآورد مقادیر $\hat{\mu}_{i1k}$ ، $\hat{\mu}_{i2k}$ ، $\hat{\sigma}_{i1k}$ ، $\hat{\sigma}_{i1k}$ و $\hat{\rho}_{12k}$ که به‌ازای آن رابطهٔ (۶) ماکسیمم می‌شود، عبارت است از [۱۴]:

$$\begin{aligned}\hat{\mu}_{i1k} &= \frac{1}{m}\sum_{j=1}^{m} a_{ij}\\ \hat{\mu}_{i2k} &= \frac{1}{m}\sum_{j=1}^{m} b_{ij}\\ \hat{\sigma}_{i1k}^2 &= \frac{1}{m-1}\sum_{j=1}^{m}(a_{ij}-\hat{\mu}_{i1k})^2\\ \hat{\sigma}_{i2k}^2 &= \frac{1}{m-1}\sum_{j=1}^{m}(b_{ij}-\hat{\mu}_{i2k})^2\\ \rho_{12k} &= \frac{1}{m\hat{\sigma}_{i1k}\hat{\sigma}_{i2k}}\sum_{j=1}^{m}(a_{ij}-\hat{\mu}_{i1k})(b_{ij}-\hat{\mu}_{i2k})\end{aligned}$$

پس از برآورد میانگین و واریانس مقادیر کران پایین و بالای صفت عددی نادقیق به روش ماکسیمم درست‌نمایی و جایگذاری آن در توزیع نرمال دومتغیره، احتمال قرارگرفتن صفت عددی نادقیق A_i در رده C_k به دست می‌آید. در این مقاله برای رده‌بندی انواع داده‌های دقیق و نادقیق، نمونه‌های آموزشی که از توزیع نرمال تبعیت می‌کنند، به الگوریتم رده‌بندی داده می‌شوند و با استفاده از مدل به دست‌آمده، قرارگرفتن نمونهٔ جدید در ردهٔ مناسب پیش‌بینی می‌شود (شکل ۱).



شکل ۱. نحوهٔ رده‌بندی نمونهٔ جدید

اکنون فرض کنید به تعداد $k = 1, 2, ..., p$ فایل داده‌های آموزشی مفروض است که هر کدام دارای نمونه‌هایی است که در یک رده قرار دارند.

۱.۴ نحوهٔ انتخاب فایل داده‌های آموزشی

۱. در ابتدا از توزیع یکنواخت در بازهٔ (۱و۰)، تعداد n مشاهده اختیار می‌کنیم، به‌طوری که

$$P(D_k) = \frac{n_k}{n_1 + n_2 + \cdots + n_p} = \frac{n_k}{n}, \quad k = 1, ..., p$$

۲. اعداد تولید شده را از کوچک به بزرگ مرتب می‌کنیم، به‌طوری که

$$\frac{n_1}{n} = \hat{n}_1 \ , \ \frac{n_2}{n} = \hat{n}_2 \ , ..., \ \frac{n_p}{n} = \hat{n}_p$$

$$\sum_{k=1}^{p} \hat{n}_k = 1$$

۳. انتخاب فایل داده‌های آموزشی بر اساس گروهی است که بیشترین تعداد مشاهده را داشته باشد، یعنی

$$D_k = \max(\hat{n}_1, ..., \hat{n}_p).$$

هریک از مقادیر نمونهٔ جدید به‌عنوان ورودی به الگوریتم رده‌بندی داده می‌شود و خروجی آن، قرار گرفتن نمونهٔ جدید در ردهٔ مناسب است. نکته مهم این است که فرض نرمال بودن داده‌های آموزشی باید برقرار باشد. لذا در بخش (۱.۵) به بررسی نرمال بودن داده‌ها می‌پردازیم.

۵ الگوریتم رده‌بندی

- گام اول: فایل داده‌های آموزشی (D_k) را به‌ازای k انتخاب کنید.

$$k = 1, 2, ..., p$$

- گام دوم: تعداد نمونه‌های D_k در m قرار داده می‌شود.
- گام سوم: هر صفت A_i از D_k را به‌ازاء $i = 1, \ldots, n$ در نظر بگیرید.
- گام چهارم:
  - اگر A_i صفت عددی دقیق باشد،

۱. مقادیر ستون A_i در a_{ij} قرار داده می‌شود.

۲.

$$\hat{\mu}_{ik} = \frac{1}{m} \sum_{j=1}^{m} a_{ij}$$

۳.

$$\hat{\sigma}^2_{ik} = \frac{1}{m-1} \sum_{j=1}^{m} (a_{ij} - \mu_{ik})^2$$

۴. احتمال شرطی صفت عددی دقیق A_i با توزیع نرمال حاصل می‌شود:

$$Pc_{ik} = N(\mu_{ik}, \sigma_{ik}).$$

  - اگر A_i صفت عددی نادقیق باشد،

۱.

$$\hat{\mu}_{i1} = \frac{1}{m} \sum_{j=1}^{m} a_{ij}$$

۲.

$$\hat{\mu}_{i2} = \frac{1}{m} \sum_{j=1}^{m} b_{ij}$$

۳.

$$\hat{\sigma}^2_{i1} = \frac{1}{m-1} \sum_{j=1}^{m} (a_{ij} - \hat{\mu}_{i1})^2$$

۴.

$$\hat{\sigma}^2_{i2} = \frac{1}{m-1} \sum_{j=1}^{m} (b_{ij} - \hat{\mu}_{i2})^2$$

۵.

$$\rho_{i12} = \frac{1}{m\hat{\sigma}_{i1}\hat{\sigma}_{i2}} \sum_{j=1}^{m} (a_{ij} - \hat{\mu}_{i1})(b_{ij} - \hat{\mu}_{i2})$$

۶.

$$\theta^*_{ik} = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \end{bmatrix}$$

۷.

$$\Sigma^*_{ik} = \begin{bmatrix} \sigma^2_{i1} & \sigma_{i1}\sigma_{i2}\rho_{i12} \\ \sigma_{i1}\sigma_{i2}\rho_{i12} & \sigma^2_{i2} \end{bmatrix}$$

۸. احتمال شرطی صفت عددی نادقیق A_i با توزیع نرمال دومتغیرهٔ زیر به دست می‌آید:

$$Pu_{ik} = N_2(\theta^*_{ik}, \Sigma^*_{ik})$$

- گام پنجم: $i \leftarrow i + 1$ و تا زمانی‌که $i \leq n$ باشد به گام سوم بر می‌گردد.
- گام ششم:

$$P(D_k) = \frac{\text{تعداد نمونه‌های } D_k}{\text{تعداد کل نمونه‌ها}}$$

- گام هفتم:

$$Pf_k = P(D_k) \times \prod_{i=1}^{n} Pu_{ik} \times Pc_{ik}$$

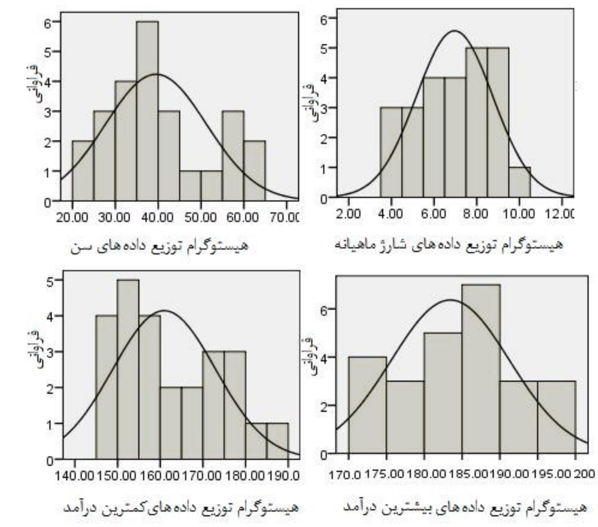
- گام هشتم: $k \leftarrow k + 1$ و تا زمانی‌که $k \leq p$ باشد، به گام اول بر می‌گردد.
- گام نهم: رده‌ای که بیشترین Pf_k را دارد، به‌عنوان خروجی قرار داده می‌شود.

۱.۵ بررسی نرمال بودن داده‌ها

در اغلب آزمون‌های پارامتری، مفروضات مقدماتی بسیاری وجود دارند که تا این مفروضات برآورده نشوند، نتایج به دست‌آمده از آزمون، نامعتبر خواهد بود. مهمترین این مفروضات، فرض نرمال بودن داده‌ها است. منظور از نرمال بودن توزیع داده‌ها این است که بافت‌نگاشت فراوانی داده‌ها تقریباً به‌صورت منحنی نرمال باشد. نمودار بافت‌نگاشت فراوانی نمونه‌های آموزشی ردهٔ «پرداخت کرده» (جدول ۳)، در شکل ۲ نشان داده شده است.

ضریب چولگی و ضریب کشیدگی دو شاخص اساسی توزیع داده‌ها هستند که با داشتن آنها تا حدی می‌توان به نرمال بودن داده‌ها پی برد.

چولگی معیاری از تقارن یا عدم تقارن تابع توزیع است. در یک توزیع کاملاً متقارن، چولگی صفر است. برای یک توزیع نامتقارن، ضریب چولگی مثبت یا منفی است. هرچه مقدار چولگی از صفر بیشتر فاصله داشته باشد، عدم تقارن، شدیدتر است.



شکل۲. بافت‌نگاشت فراوانی صفت‌های «رده پرداخت کرده»

کشیدگی نشان‌دهندهٔ ارتفاع یک توزیع است. به عبارت دیگر کشیدگی معیاری از بلندی منحنی در نقطهٔ ماکسیمم است. همیشه کشیدگی را با کشیدگی توزیع نرمال مقایسه می‌کنند. مقدار کشیدگی برای توزیع نرمال برابر با ۳ است.

در حالت کلی چنانچه مقدار چولگی و کشیدگی داده‌ها خارج از بازهٔ (۲و۲−) باشد، داده‌ها از توزیع نرمال برخوردار نیستند و باید قبل از هر آزمونی که با نتیجهٔ آن با فرض نرمال بودن داده‌ها برقرار است، آن را به توزیع نرمال نزدیک‌تر کرد. وقتی این دو شاخص در حالت نرمال قرار نداشته باشند، می‌توان حدس زد که توزیع داده‌ها نرمال نیست. برای اثبات این ادعا می‌توان از آزمون‌های آماری استفاده کرد. هنگام بررسی نرمال بودن داده‌ها فرض صفر را مبنی بر این‌که توزیع داده‌ها نرمال است، در نظر می‌گیریم و همانند اکثر آزمون‌ها، آن را در سطح خطای ۵ درصد بررسی می‌کنیم. با توجه به این‌که نمونه‌های آموزشی این مقاله کمتر از ۲۰۰۰ نمونه است، بررسی نرمال بودن داده‌ها را با آزمون شاپیرو–ویلک[۹] در نرم‌افزار SPSS انجام داده‌ایم [۲، ۳].

[۹] Shapiro-Wilk

۶ نتایج عددی

نحوهٔ یادگیری مدل ارائه‌شده، بر اساس یادگیری باناظر است؛ یعنی با استفاده از نمونه‌های آموزشی در ابتدا مدل رده‌بندی آموزش داده می‌شود و سپس بر اساس این آموزش‌ها می‌تواند نمونهٔ جدید را در رده مناسب قرار دهد. در این مقاله ۵۰ نمونهٔ مختلف از بازپرداخت اقساط افرادی که در دو رده (پرداخت کرده. پرداخت نکرده) قرار دارند، به‌عنوان نمونه‌های آموزشی برای یادگیری مدل رده‌بندی استفاده شده است که در آن سن و شارژ ماهانه از جمله صفت‌های عددی دقیق و درآمد ماهانه به‌عنوان صفت عددی نادقیق در نظر گرفته شده است (جدول ۳). نرمال بودن داده‌های مربوط به هر صفت عددی دقیق یا نادقیق که در یک رده قرار دارند، با نرم‌افزار SPSS بررسی و سپس نمونه‌های هر رده در یک دادگان از نرم‌افزار Matlab ذخیره می‌شود. با فراخوانی هر یک از این دادگان‌ها توسط الگوریتم رده‌بندی ارائه‌شده، نمونهٔ جدید در ردهٔ مناسب قرار می‌گیرد. به‌عنوان مثال، نمونه‌های جدول ۱، نمونه‌های جدیدی هستند که پس از برآورد پارامترهای مربوط به هر یک از صفت‌های عددی دقیق و نادقیق در ردهٔ مناسب قرار گرفته‌اند نمونه‌های ۵۲ و ۵۴ در ردهٔ افرادی رده‌بندی شده‌اند که اقساط خود را پرداخت کرده‌اند و نمونه‌های ۵۱ و ۵۳ در رده افرادی که اقساط خود را پرداخت نکرده‌اند. برآورد پارامترهای میانگین و واریانس هر یک از صفت‌های عددی دقیق و نادقیق مربوط به رده‌های «پرداخت نکرده» و «پرداخت کرده» به‌ترتیب در جدول‌های ۴ و ۵ آورده شده است.

۱.۶ ارزیابی مدل آمیختهٔ بیزی

مهمترین معیار ارزیابی کارایی یک الگوریتم رده‌بندی، دقت یا نرخ صحت[۱۰] رده‌بندی است که نشان می‌دهد مدل طراحی‌شده چند درصد از کل نمونه‌های آزمایشی را به‌درستی رده‌بندی کرده است. برای به دست آوردن دقت مدل، در ابتدا باید با مفهوم ماتریس درهم‌ریختگی[۱۱] آشنا شویم. ماتریس درهم‌ریختگی، یک ماتریس $n \times n$ است که چگونگی عملکرد یک الگوریتم رده‌بندی را با توجه به تعداد رده‌بندی درست و نادرست انجام شده توسط مدل رده‌بندی در مقایسه با نتایج واقعی نشان می‌دهد و معمولاً برای الگوریتم‌های یادگیری باناظر استفاده می‌شود. ماتریس درهم‌ریختگی برای دو رده ($n = ۲$) در شکل ۳ نشان داده شده است که شامل ۴ عنصر TP، TN، FP و FN است [۴]:

TP، تعداد نمونه‌هایی که ردهٔ واقعی آن‌ها «پرداخت کرده» بوده و به‌درستی در ردهٔ «پرداخت کرده» رده‌بندی شده‌اند.

TN، تعداد نمونه‌هایی که ردهٔ واقعی آن‌ها «پرداخت نکرده» بوده و به‌درستی در ردهٔ «پرداخت نکرده» رده‌بندی شده‌اند.

FP، تعداد نمونه‌هایی که ردهٔ واقعی آن‌ها «پرداخت نکرده» بوده و به‌اشتباه در ردهٔ «پرداخت کرده» رده‌بندی شده‌اند.

FN، تعداد نمونه‌هایی که ردهٔ واقعی آن‌ها «پرداخت کرده» بوده و به‌اشتباه در ردهٔ «پرداخت نکرده» رده‌بندی شده‌اند.

جدول ۱. نمونه‌های جدید

نمونه	سن	شارژ ماهانه (دلار)	درآمد ماهانه (دلار)
۵۱	۴۲	۹	[۶۵ و ۹۰]
۵۲	۲۶	۶	[۱۶۰ و ۱۸۰]
۵۳	۳۱	۵	[۹۰ و ۱۰۰]
۵۴	۵۰	۸	[۱۵۵ و ۱۶۵]

حساسیت،[۱۲] ویژگی[۱۳] و دقت[۱۴] از جمله معیارهای عملکرد مسائل رده‌بندی هستند. این معیارها به‌طور خلاصه در ادامه شرح داده شده‌اند [۱].

ماتریس در هم ریختگی		رده تشخیص داده شده	
		پرداخت کرده	پرداخت نکرده
رده واقعی	پرداخت کرده	TP	FN
	پرداخت نکرده	FP	TN

شکل ۳. ماتریس درهم‌ریختگی برای دو رده

حساسیت:

حساسیت نشان می‌دهد اگر نتیجهٔ واقعی برای نمونهٔ جدید «پرداخت کرده» باشد، در چند درصد موارد مدل نیز نتیجهٔ

[۱۰] accuracy rate

[۱۱] confusion matrix

[۱۲] sensitivity

[۱۳] specificity

[۱۴] accuracy

«پرداخت کرده» خواهد داشت، که از رابطهٔ زیر حساب می‌شود:

$$\text{حساسیت} = \frac{TP}{TP+FN} \tag{۷}$$

ویژگی:

ویژگی نشان می‌دهد اگر نتیجهٔ واقعی برای نمونهٔ جدید «پرداخت نکرده» باشد، در چند درصد موارد مدل نیز نتیجهٔ «پرداخت نکرده» خواهد داشت، که از رابطهٔ زیر به دست می‌آید:

$$\text{ویژگی} = \frac{TN}{FP+TN} \tag{۸}$$

دقت:

دقت به معنای درصد رده‌های درست پیش‌بینی‌شده است. در این معیار، دو مقدار TP و TN مهم‌ترین مقادیری هستند که در یک مسئلهٔ دو رده‌ای باید بیشینه شوند و از رابطهٔ زیر حساب می‌شود:

$$\text{دقت} = \frac{TN+TP}{TN+FN+TP+FP} \tag{۹}$$

به‌منظور ارزیابی کارایی مدل رده‌بندی ارائه‌شده و مدل کین و همکاران [۱۳]، نمونه‌های آموزشی را یک‌بار دیگر به‌عنوان نمونه‌های آزمایشی مورد بررسی قرار داده‌ایم، که هر دو مدل نتایج درستی را نشان داده‌اند. علاوه برآن، ۷۰ نمونهٔ آزمایشی جدید را که خروجی درست آن را می‌دانستیم، با نتایج حاصل از هر دو مدل رده‌بندی مورد بررسی قرار داده‌ایم، که نتایج آن در جدول ۲ نشان داده شده است.

بحث و نتیجه‌گیری

تا کنون رده‌بندی داده‌های دقیق با روش‌های مختلفی از قبیل رگرسیون، خوشه‌بندی، درخت تصمیم، بیزی و غیره مورد بررسی و تحلیل قرار گرفته است. از میان این روش‌ها، رده‌بندی سادهٔ بیزی، یادگیری را به‌خوبی انجام می‌دهد و نیاز به برآورد تکراری پارامترها ندارد و می‌تواند حجم عظیمی از داده‌ها را رده‌بندی کند [۱۷]. بنا بر این در این مقاله برای رده‌بندی داده‌های دقیق از رده‌بندی سادهٔ بیزی استفاده نموده‌ایم. از آن‌جا که داده‌ها ممکن است متعلق به بازه‌ای از مقادیر باشند، رده‌بندی نادقیق مطرح شده است. برآوردهای مختلفی برای میانگین و واریانس توزیع حاکم بر داده‌های نادقیق با فرض نرمال بودن توزیع آن، بر اساس یادگیری باناظر انجام شده است [۱۲، ۱۳]. در این مقاله فرض کرده‌ایم توزیع داده‌های نادقیق، توزیع نرمال دومتغیره است. بر روی هر یک از مقادیر دو سر بازهٔ داده‌های نادقیق، برآوردهای میانگین و واریانس را با روش ماکسیمم درست‌نمایی به دست آورده‌ایم. سپس بر اساس رده‌بندی سادهٔ بیزی، یک مدل آمیختهٔ بیزی برای رده‌بندی داده‌های دقیق و نادقیق ارائه کرده‌ایم که میزان ویژگی، حساسیت و دقت به‌دست‌آمده از آن، می‌تواند این روش را در کنار روش‌های رده‌بندی داده‌های دقیق و نادقیق، به‌عنوان الگوریتمی کم‌هزینه و کاربردی قرار دهد.

جدول ۲. دقت، حساسیت، ویژگی در مدل بیزی کین و همکاران و مدل آمیختهٔ بیزی

	مدل آمیختهٔ بیزی	مدل کین و همکاران
TP	۳۴	۳۴
TN	۳۴	۳۴
FP	۱	۱
FN	۱	۱
حساسیت	۰.۹۷	۰.۹۷
ویژگی	۰.۹۷	۰.۹۷
دقت	۰.۹۷	۰.۹۷

جدول ۳. نمونه‌های آموزشی

	اقساط خود را پرداخت نکرده‌اند		
نمونه	سن	شارژ ماهانه	درآمد ماهانه
۱	۴۵	۱۰	[۶۰ و ۱۰۰]
۲	۳۵	۸	[۷۵ و ۹۸]
۳	۲۸	۷	[۸۸ و ۹۳]
۴	۲۶	۷	[۸۵ و ۱۰۰]
۵	۵۶	۸	[۹۵ و ۱۰۰]
۶	۵۸	۹	[۶۵ و ۹۲]
۷	۳۹	۱۰	[۷۹ و ۱۱۰]
۸	۴۲	۶	[۸۶ و ۹۴]
۹	۵۰	۷	[۹۷ و ۱۱۰]
۱۰	۶۰	۹	[۶۶ و ۹۸]
۱۱	۳۶	۸	[۸۳ و ۱۰۰]
۱۲	۴۰	۹	[۹۰ و ۹۹]
۱۳	۵۳	۱۰	[۹۴ و ۱۰۰]
۱۴	۲۵	۴	[۸۰ و ۱۰۰]
۱۵	۳۰	۴	[۹۰ و ۱۰۰]
۱۶	۴۰	۸	[۸۷ و ۹۰]
۱۷	۲۵	۵	[۸۵ و ۹۵]
۱۸	۵۰	۵	[۱۱۵ و ۱۲۰]
۱۹	۳۳	۶	[۱۲۰ و ۱۳۰]
۲۰	۲۹	۷	[۸۰ و ۹۵]
۲۱	۳۴	۸	[۷۰ و ۸۵]
۲۲	۲۷	۵	[۶۵ و ۷۵]
۲۳	۳۵	۶	[۹۰ و ۱۱۵]
۲۴	۴۹	۶	[۶۵ و ۸۵]
۲۵	۳۰	۴	[۹۶ و ۱۰۵]

	اقساط خود را پرداخت کرده‌اند		
نمونه	سن	شارژ ماهانه	درآمد ماهانه
۲۶	۳۲	۵	[۱۴۵ و ۱۷۰]
۲۷	۳۸	۴	[۱۴۷ و ۱۷۵]
۲۸	۲۵	۷	[۱۵۸ و ۱۸۰]
۲۹	۲۹	۶	[۱۴۸ و ۱۸۵]
۳۰	۲۲	۸	[۱۵۱ و ۱۹۰]
۳۱	۳۵	۶	[۱۵۶ و ۱۸۸]
۳۲	۴۰	۴	[۱۵۰ و ۱۷۹]
۳۳	۲۷	۴	[۱۴۷ و ۱۸۷]
۳۴	۳۰	۶	[۱۵۵ و ۱۷۴]
۳۵	۳۴	۵	[۱۵۰ و ۱۷۷]
۳۶	۲۳	۶	[۱۵۰ و ۱۸۹]
۳۷	۳۳	۷	[۱۵۵ و ۱۸۵]
۳۸	۵۰	۸	[۱۶۰ و ۱۷۰]
۳۹	۳۵	۷	[۱۷۰ و ۱۸۵]
۴۰	۴۲	۹	[۱۷۲ و ۱۹۰]
۴۱	۶۱	۸	[۱۶۵ و ۱۸۹]
۴۲	۴۳	۹	[۱۷۹ و ۱۹۰]
۴۳	۴۸	۹	[۱۸۰ و ۱۹۵]
۴۴	۵۵	۸	[۱۷۷ و ۱۸۰]
۴۵	۶۱	۹	[۱۷۰ و ۱۹۸]
۴۶	۵۵	۱۰	[۱۶۵ و ۱۸۰]
۴۷	۳۷	۹	[۱۷۵ و ۱۸۳]
۴۸	۳۶	۷	[۱۸۵ و ۱۹۵]
۴۹	۵۹	۸	[۱۶۰ و ۱۸۰]
۵۰	۳۷	۵	[۱۵۳ و ۱۷۲]

جدول ۴. تخمین پارامترهای نمونه‌های جدید در ردهٔ پرداخت نکرده

	نمونهٔ ۵۱	نمونهٔ ۵۲	نمونهٔ ۵۳	نمونهٔ ۵۴
میانگین سن	۳۹	۳۹	۳۹	۳۹
انحراف معیار سن	۱۱۹/۴۱۶۷	۱۱۹/۴۱۶۷	۱۱۹/۴۱۶۷	۱۱۹/۴۱۶۷
میانگین شارژ	۷/۰۴۰۰	۷/۰۴۰۰	۷/۰۴۰۰	۷/۰۴۰۰
انحراف معیار شارژ	۳/۶۲۳۳	۳/۶۲۳۳	۳/۶۲۳۳	۳/۶۲۳۳
میانگین کمترین درآمد	۸۴/۲۴۰۰	۸۴/۲۴۰۰	۸۴/۲۴۰۰	۸۴/۲۴۰۰
انحراف معیار کمترین درآمد	۱۴/۷۸۵۴	۱۴/۷۸۵۴	۱۴/۷۸۵۴	۱۴/۷۸۵۴
میانگین بیشترین درآمد	۹۸/۴۸۰۰	۹۸/۴۸۰۰	۹۸/۴۸۰۰	۹۸/۴۸۰۰
انحراف معیار بیشترین درآمد	۱۱/۴۰۹۵	۱۱/۴۰۹۵	۱۱.۴۰۹۵	۱۱/۴۰۹۵

جدول ۵. تخمین پارامترهای نمونه‌های جدید در ردهٔ پرداخت کرده

	نمونهٔ ۵۱	نمونهٔ ۵۲	نمونهٔ ۵۳	نمونهٔ ۵۴
میانگین سن	۳۹/۴۸۰۰	۳۹/۴۸۰۰	۳۹/۴۸۰۰	۳۹/۴۸۰۰
انحراف معیار سن	۱۳۸/۶۷۶۷	۱۳۸/۶۷۶۷	۱۳۸/۶۷۶۷	۱۳۸/۶۷۶۷
میانگین شارژ	۶/۹۶۰۰	۶/۹۶۰۰	۶/۹۶۰۰	۶/۹۶۰۰
انحراف معیار شارژ	۳/۲۰۶۷	۳/۲۰۶۷	۳/۲۰۶۷	۳/۲۰۶۷
میانگین کمترین درآمد	۱۶۰/۹۲۰۰	۱۶۰/۹۲۰۰	۱۶۰/۹۲۰۰	۱۶۰/۹۲۰۰
انحراف معیار کمترین درآمد	۱۲/۰۴۱۳	۱۲/۰۴۱۳	۱۲/۰۴۱۳	۱۲/۰۴۱۳
میانگین بیشترین درآمد	۱۸۳/۴۴۰۰	۱۸۳/۴۴۰۰	۱۸۳/۴۴۰۰	۱۸۳/۴۴۰۰
انحراف معیار بیشترین درآمد	۷/۸۳۲۰	۷/۸۳۲۰	۷/۸۳۲۰	۷/۸۳۲۰

مراجع

[۱] باقرزاده خیابانی و اخوان نیاکی (۱۳۹۲). ارائهٔ یک مدل ترکیبی به جهت پیش‌بینی بیماری دیابت نوع ۲. هفتمین کنفرانس داده‌کاوی ایران.

[2] Marques, D. JP. (2003). *Applied Statistics: Using SPSS, STATISTICA, and MATLAB*, Springer Science and Business Media.

[3] Surhone, L. M., Timpledon, M.T., and Marseken, S.F. (2010). *Shapiro-Wilk Test*, VDM Publishing.

[4] Alpaydin, E. (2004). *Introduction to Machine Learning*, MIT Press.

[5] Billard, L. and Diday, E. (2000). Regression analysis for interval-valued data, *Data Analysis: Classification and Related Methods*, Springer, 369-374.

[6] Carvalho, F., Lima Neto, E., and Tenorio, C. (2004). A new method to fit a linear regression model for interval valued data, *Advances in Artificial Intelligence*, Springer, 295-306.

[7] Carvalho, F., Brito, P., and Bock, H. H. (2006). Dynamic clustering for interval data based on L2 distance, *Computational Statistics*, Springer , **21** ,231–250.

[8] Chavent, M., Carvalho, F., Lechevallier, Y., and Verde, R. (2006). New clustering methods for interval data, *Computational Statistics*, Springer, **21** , 211–229.

[9] Cormode, G. and McGregor, A. (2008). Approximation algorithm for clustering uncertain data, *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 191–199.

[10] Kriegel, H. and Pfeifle, M. (2005). Density based clustering of uncertain data, *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* , 672–677.

[11] Qin, B., Xia, Y., and Li, F. (2009). DTU: a decision tree for uncertain data, *Advances in Knowledge Discovery and Data Mining* , Springer, 4–15.

[12] Qin, B., Xia,Y., and Li, F. (2010). A Bayesian classifier for uncertain data, *Proceedings of the 2010 ACM Symposium on Applied Computing*, 1010–1014.

[13] Qin, B. , Xia, Y. , Wang, S., and Du,X. (2011). A novel Bayesian classification for uncertain data, *Knowledge-Based Systems*, Elsevier, **24**, 1151-1158.

[14] Le, R. and Billard, L. (2011). Likelihood functions and some maximum likelihood estimators for symbolic data, *Statistical Planning and Inference*, **141**, 1593-1602.

[15] Wackerly, D. Mendenhall, W., and Scheaffer, R. (2007). *Mathematical Statistics with Applications*, Cengage Learning.

[16] Mitchell, T. (1997). *Machine Learning*, McGraw-Hill science.

[17] Wu, X. , Kumar, V., Quinlan, J. R., and Motoda, H. (2008). Top 10 algorithms in data mining, *Knowledge and Information Systems*, Springer ,**14** ,1-37.