dataacademy.ir

رگرسیون درختی و طبقه‌بندی

مجتبی کاظمی[۱]، عظیم بازرگانی گیلانی[۲]
[۱] کارشناس ارشد عمران، باشگاه پژوهشگران جوان و نخبگان دانشگاه آزاد اسلامی، واحد رودسر و املش، رودسر، گیلان.
mojtaba.kazemi88@yahoo.com
[۲] کارشناس عمران، معاونت عمرانی، دانشگاه آزاد اسلامی، واحد رودسر و املش، رودسر، گیلان.
Azim.bazargani@gmail.com

چکیده – داده کاوی یک علم میان رشته‌ای با هدف استخراج دانش پنهان از میان بانک اطلاعاتی انبوه می‌باشد. رگرسیون درختی و طبقه‌بندی از روش‌های مهم داده‌کاوی است و مدلی ناپارمتریک و بدون پیش فرض خاص محسوب می‌شود. در این روش شاخه‌های دوتایی بر اساس یک متغیر مستقل ایجاد می‌شوند. معیار ارزیابی شاخه‌ها، گوناگونی نام دارد. برای جداسازی گره به دو زیر گره می‌توان از شاخص جینی یا شاخص دوتایی استفاده نمود. مهمترین و اصلی‌ترین معیار ارزیابی درخت ایجاد شده، معیار نرخ خطا در درخت است. به منظور محاسبه‌ی نرخ خطای کل درخت، مجموع وزنی نرخ خطاهای برگ‌ها بدست آورده می‌شود. به منظور جلوگیری از تولید قانون‌های بی‌کیفیت در برخی از شاخه‌ها، هرس صورت می‌گیرد. هرچند این عمل باعث افزایش نرخخطا می‌شود، اما مانع از ایجاد برخی قانون‌های ناکارا می‌شود. توجه به این نکته نیز ضروری است که هرس به نحوی صورت گیرد تا خطا از مقدار معینی بیشتر نشود. در نهایت باید توجه داشت که درختی بهینه است که کمترین هزینه‌ی دسته‌بندی اشتباه را برای داده‌های آزمایشی داشته باشد.

کلید واژه- داده‌کاوی، رگرسیون درختی و طبقه‌بندی، نرخ خطا، شاخص جینی، هرس.

۱- مقدمه

داده‌کاوی یک علم میان رشته‌ای است که روش‌های مختلف از جمله آمار، تشخیص الگو، یادگیری ماشین و پایگاه داده را بکار می‌گیرد تا دانش نهفته در انبوه داده‌های بانک اطلاعاتی را استخراج کند. در دسترس بودن حجم وسیعی از داده‌ها و نیاز شدید به اینکه از این داده‌ها اطلاعات و دانش سودمندی استخراج کنیم، باعث شد داده‌کاوی کانون توجهات در صنعت اطلاعات قرار بگیرد. اطلاعات و دانش بدست آمده در کاربردهای وسیعی از مدیریت کسب و کار، کنترل تولید و تحلیل بازار تا طراحی مهندسی و تحقیقات علمی استفاده می‌شود [۱].

مدل‌های درختی که در رده‌بندی و رگرسیون درختی استفاده می‌شوند توسط مورگان و سونی کوئیست در سال ۱۹۶۳ میلادی برای بررسی اثرات متقابل متغیرها در داده‌های علوم اجتماعی پیشنهاد شده‌اند و جنبه‌های نظری و کاربردی آن توسط بریمن و همکارانش در سال ۱۹۸۴ میلادی در رساله‌ای که در این مورد منتشر گردید، بسط و توسعه داده شد [۲].

مدل کارت که یک مدل ناپارامتری و بدون هرگونه پیش فرض در خصوص رابطه بین متغیرهای مستقل و متغیر هدف است و از روش‌های مهم داده‌کاوی می‌باشد، به طور گسترده در تجارت، صنعت، مهندسی و سایر علوم استفاده شده است. مدل کارت، ابزاری قدرتمند در تعیین مهم‌ترین متغیرهای مستقل و حل مسائل دسته‌بندی و پیش‌بینی است [۳].

به طور کلّی روش‌های مبتنی بر مدل‌های خطی، فضای متغیرهای کمکی را به ناحیه‌های مجزا تقسیم می‌کند و داده‌ها را به گروه‌های متناظر تخصیص می‌دهد. این روش‌ها داده‌ها را به طور بازگشتی برای تعیین یا معرفی اثرات متقابل متغیرها و معرفی زیر گروه‌هایی از افراد با مشخصات دموگرافی و علائم مشخص جهت تشخیص‌های بعدی تقسیم می‌کند. با توجه به نوع مسئله، هدف اساسی در یک مطالعه‌ی مدل‌های رده‌بندی و رگرسیون درختی می‌تواند ایجاد یک رده‌بندی کننده‌ی دقیق و یا کشف یک ساختار پیش‌بینی کننده برای مسئله مورد نظر باشد. اگر هدف تعیین یک ساختار پیش‌بینی کننده باشد، آنگاه درک صحیح متغیرها و اثرات متقابل آنها ضروری است. معمولاً

dataacademy.ir

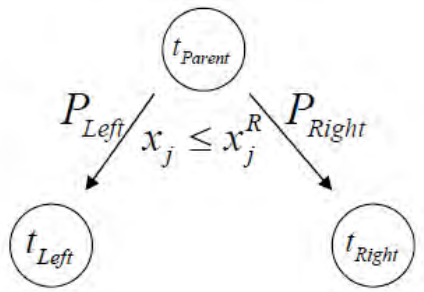
در مسائل مختلف این دو هدف به موازات هم بررسی می‌شوند [۲].

۲- درخت تصمیم

طبقه‌بندی و رگرسیون دو مسأله مهم در علم آمار می‌باشند. الگوریتم طبقه‌بندی و درخت رگرسیون در برگیرنده سه وظیفه مهم است. اولین وظیفه این است که چگونه در هر مرحله داده‌ها را بخش‌بندی نماید. دومین وظیفه آن است که چه زمانی بخش-بندی را متوقف نماید. آخرین وظیفه، چگونگی پیش‌بینی مقدار y برای هر x در یک بخش‌بندی (قسمت)، است [۴]. همچنین این روش به سه دلیل عمده، نوع جذاب و مخصوص مدل‌ها می-باشد. اول آن‌که نشان‌دهنده نتایج مدل به صورت آسان برای درک و تلفیق (شبیه‌سازی) توسط انسان می‌باشد. دومین دلیل این است که درخت تصمیم مدلی ناپارامتریک می‌باشد، مداخله توسط کاربر نیاز ندارد، و بسیار مناسب برای جستجوی دانش اکتشافی می‌باشد. سوم این که الگوریتم قابل درجه‌بندی است، به مفهوم دیگر؛ کارایی درجه‌بندی مطلوب ارتباط با افزایش اندازه نمونه آموزشی دارد، این حالت برای درخت تصمیم مدل‌های ساخته شده وجود دارد و نیز صحت درخت تصمیم، همسنگ یا برتر از دیگر مدل‌ها می‌باشد [۵].

۳- درخت طبقه‌بندی

درخت طبقه‌بندی زمانی برای مشاهده نمونه آموزشی به کار برده می‌شود که بدانند طبقه در حال پیشرفت (حالت پیشرفته) است. طبقات در نمونه آموزشی شاید توسط کاربر فراهم شده باشد یا در مطابقت با برخی قوانین برون‌زا محاسبه گردد. در شکل ۱، t_p به عنوان یک گره والدین (مادر) و t_r و t_l به ترتیب گره‌های فرزند راست و چپ گره والدین می‌باشد. ملاحظه می-شود که نمونه آموزشی با ماتریس متغیر X با M تعداد متغیرهای x_j و N مشاهدات می‌باشد. در اینجا بردار طبقه Y شامل N مشاهدات با مجموع مقدار K طبقه می‌باشد. درخت طبقه‌بندی در تطابق با قانون جداسازی ساخته می‌شود. قانونی که ایفاگر جداسازی نمونه آموزشی به بخش‌های کوچک‌تر است.



شکل (۱): الگوریتم جداسازی CART

که در اینجا؛ t_p، t_r و t_l گره‌های والدین، راست و چپ می‌باشند؛ x_j متغیر j؛ x_j^R بهترین مقدار جدا شده متغیر x_j بیشترین (حداکثر) همجنسی گره‌های فرزند با نام تابع ناپاکی i(t) تعریف شده است. از این جهت ناپاکی گره والدین t_p برای هر جداسازی ممکن $x_j \le x_j^R, j = ۱,۲,\dots,M$ ثابت است. حداکثر یکنواختی گره‌های فرزند چپ و راست برابر بیشینه تغییر تابع ناپاکی $\Delta i(t)$ خواهد بود:

$$\Delta i(t) = i(t_p) - E[i(t_c)] \qquad (۱)$$

که در اینجا t_c گره‌های فرزند چپ و راست گره والدین t_p هستند. فرض بر آن است که P_l و P_r احتمال گره‌های چپ و راست است. بنابراین بدست می‌آید:

$$\Delta i(t) = i(t_p) - P_l i(t_l) - P_r i(t_r) \qquad (۲)$$

در نتیجه، در هر گره CART، مشکل بیشینه ذیل حل می-گردد:

$$\underset{x_j \le x_j^R, j = ۱,۲,\dots,M}{\arg\max} \; [i(t_p) - P_l i(t_l) - P_r(t_r)] \qquad (۳)$$

رابطه (۳) این مفهوم را می‌رساند که CART به طور کامل تمام مقادیر ممکن کل متغیرها در ماتریس X برای بهترین جداسازی پرسش $x_j \le x_j^R$ که بیشینه تغییر سنجش ناپاکی $\Delta i(t)$ خواهد بود را مورد تجسس قرار می‌دهد.

سئوال مهم بعدی در مورد چگونگی تعریف تابع ناپاکی i(t) خواهد بود. در این تئوری توابع ناپاکی متعدد بوده‌اند. دو گونه بسیار مهم و کاربردی عبارتند از؛ قاعده جداسازی جینی و قاعده جداسازی توئینگ (دوتایی).

۳-۱- قاعده جداسازی جینی

قاعده جداسازی جینی (یا شاخص جینی) قانونی می‌باشد که به نسبت وسیعی مورد استفاده قرار می‌گیرد. این کاربرد تابع ناپاکی i(t) طبق رابطه ۴ می‌باشد:

$$i(t) = \sum_{k \ne ۱} p(k|t)p(l|t) \qquad (۴)$$

dataacademy.ir

در اینجا k و l، ۱و.... Kو شاخص طبقه، p(k|t) احتمال شرطی طبقه k که از گره t بدست می‌آید.

بکار گرفتن تابع ناپاکی جینی بنابر بیشینه، فرمول (۴)، این امکان را فراهم می‌آورد تا تغییر سنجش ناپاکی $\Delta i(t)$ بدست آید:

$$\Delta i(t) = -\sum_{k=1}^{K} p^2(k|t_p) + P_l \sum_{k=1}^{K} p^2(k|t_l) + P_r \sum_{k=1}^{K} p^2(k|t_r)$$

بنابراین، الگوریتم جینی به صورت فرمول (۵) محاسبه می‌گردد:

(۵)

$$\underset{x_j \le x_j^R, j=1,2,\dots,M}{\arg\mathrm{Max}} \left[-\sum_{k=1}^{K} p^2(k|t_p) + P_l \sum_{k=1}^{K} p^2(k|t_l) + P_r \sum_{k=1}^{K} p^2(k|t_r)\right]$$

الگوریتم جینی در نمونه آموزشی برای بزرگترین طبقه و مجزا نمودن از سایر داده‌ها (باقیمانده‌ها) جستجو می‌نماید. جینی برای داده‌های با اغتشاش مناسب است.

۳-۲- قاعده جداسازی دوتایی

این قاعده؛ غیر مشابه با قاعده جینی بوده و برای دو طبقه که با یکدیگر بیش از ۵۰٪ داده‌ها را تشکیل می‌دهند، مورد تفحص قرار می‌دهد. قاعده جداسازی دوگانه، تغییر سنجش ناپاکی ذیل را بیشینه می‌سازد:

$$\Delta i(t) = \frac{P_l P_r}{4} \left[\sum_{k=1}^{K} |p(k|t_l) - p(k|t_r)|\right]^2 \qquad (۶)$$

که دلالت بر بیشینه مقدار فرمول (۷) دارد:

$$\left(\frac{P_l P_r}{4} \left[\sum_{k=1}^{K} |p(k|t_l) - p(k|t_r)|\right]^2\right) \qquad (۷)$$

$$\underset{x_j \le x_j^R, j=1,2,\dots,M}{\arg\mathrm{Max}}$$

همچنین قاعده جداسازی دوگانه جهت ایجاد نمودن درخت‌های متوازن بیشتر، اجازه خواهد داد. عملکرد این الگوریتم نسبت به قاعده جینی کُندتر می‌باشد. برای مثال، اگر تعداد کل طبقات برابر K باشد، بنابراین 2^{k-1} جداکننده خواهد داشت. در تحقیق کنونی از قاعده جداسازی جینی استفاده شده است.

۴- رگرسیون درختی

رگرسیون درختی قابلیت طبقه‌بندی را ندارد. در عوض این بردار پاسخ Y می‌باشد که نشان دهنده مقادیر پاسخ برای هر مشاهده در ماتریس متغیر X است. از آنجایی‌که درخت رگرسیون پیش اختصاص طبقه‌بندی انجام نمی‌دهد، قواعد جداسازی طبقه‌بندی مشابه جینی یا دوتایی کاربردی نخواهند بود. جداسازی در رگرسیون درختی مطابق با الگوریتم حداقل مربع باقیمانده با دلالت بر آنکه مجموع واریانس‌های مورد انتظار برای دو نتیجه‌گیری گره‌ها باید حداقل شده باشد، ساخته می‌شود.

(۸)

$$\underset{x_j \le x_j^R, j=1,2,\dots,M}{\arg\min} \left[P_l Var(Y_l) + P_r Var(Y_r)\right]$$

که در اینجا $Var(Y_l)$ و $Var(Y_r)$، بردارهای پاسخ برای متناظر بودن گره‌های فرزند چپ و راست می‌باشد.

$x_j \le x_j^R, j = 1,\dots,M$ بهینه پرسش‌های جداسازی که هر کدام رضایت‌مندی شرایط فرمول (۸) می‌باشد.

الگوریتم حداقل مربع باقیمانده برابر با قاعده جداسازی جینی است. تابع ناپاکی جینی (۴) در توصیف کامل نکته واریانس‌ها ساده است. اگر به مقادیر طبقه K مقدار ۱ و به مقادیر سایر طبقات عدد صفر را اختصاص دهند، بدین‌گونه واریانس نمونه این مقادیر برابر $p(k|t)[1 - p(k|t)]$ خواهد بود. در مجموع توسط شماره (تعداد) طبقات K، می‌توان معیار سنجش ناپاکی i(t) ذیل را بدست آورد:

$$i(t) = 1 - \sum_{k=1}^{K} p^2(k|t) \qquad (۹)$$

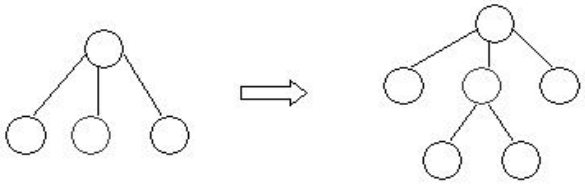
در بالا این نکته به نام درخت حداکثر ساخته شده بود و به این معنا می‌باشد که جداسازی برای مشاهدات گذشته در نمونه آموزشی ساخته شده است. درخت حداکثر شاید حاصل بسیار بزرگی باشد، بویژه در نمونه رگرسیون درختی، زمانی‌که احتمال هر مقدار پاسخ نتیجه‌ای در یک گره مجزا باشد [۶].

۴-۱- عمل شاخه بندی در رگرسیون درختی (تقسیم)

عمل شاخه‌بندی پایه ساخت درخت است. یک گره از درخت و یک خصوصیت X_j را در نظر می‌گیرند. فرض کنید بازه تعریف این خصوصیت به تعداد L_j زیر مجموعه تقسیم شود (در زیر روش‌های انتخاب چنین زیر مجموعه‌هایی خواهد آمد). در مورد خصوصیت‌های کمی، این زیر مجموعه‌ها گروهی از زیر بازه‌های مجزا هستند، در مورد خصوصیت‌های کیفی، زیر مجموعه‌ای از مقادیر و در مورد داده‌های خصوصیات ترتیبی زیر مجموعه‌هایی شامل مقادیر همسایه می‌باشند.

فرض بر آن است به هر یک از زیر مجموعه‌ها یک شاخه نسبت داده شود که از گره فعلی (مادر) به سمت یک گره جدید که (فرزند) نامیده می‌شود امتداد یابد. بنابراین، گره به تعداد L_j گره جدید "منشعب" شده (تقسیم شده) است (شکل ۲).

dataacademy.ir



شکل (۲): تقسیم در رگرسیون درختی

توجه شود برای درخت‌های باینری L_j همیشه برابر با ۲ است. اگر L_j برای یک درخت همیشه برابر با ۳ باشد آن درخت را "سه‌گانه" می‌نامند. اگر L_j همیشه برابر با ۴ باشد یک درخت "چهارگانه" خواهد بود.

برای خصوصیت‌های کیفی یا ترتیبی مواردی می‌تواند وجود داشته باشد (هنگامی‌که اندازه نمونه مشاهدات کوچک است) که مجموعه مقادیر مشاهدات خصوصیت انجام شده برای یک گره تنها قسمتی از کل بازه تعریف مقادیر خصوصیت را در بَر بگیرد. در چنین مواردی لازم است که بقیه مقادیر را به یک شاخه جدید نسبت دهند، تا در پیش‌بینی نمونه کنترل که دارای چنین مقداری باشد بتوان تعیین نمود که این مقدار به کدام شاخه تعلق دارد. برای مثال، ممکن است مقادیر معلوم را با توجه به بیشترین تعداد مشاهدات به شاخه‌ها نسبت داد.

۴-۲- عمل تعریف درجه توافق برای شاخه بندی گره (قانون توقف)

یک گره آزاد (گره‌ای که شاخه‌ای از آن منشعب نشده) را در درخت در نظر بگیرید، که مشخص نیست آیا این گره یک برگ است یا این‌که باید شاخه‌بندی شود. زیر مجموعه مشاهدات مربوط به این گره را در نظر می‌گیرند. گره‌ها را به دو دسته تقسیم می‌کنند؛ اول آنکه این مقادیر همگن باشند، یعنی اساساً متعلق به یک کلاس باشند (مسئله تشخیص الگو RP). یا اینکه واریانس Y آنها به اندازه کافی کوچک باشد (مسئله آنالیز رگرسیون RA). موردی که در آن مقدار خصوصیت برای تمام مشاهدات یکسان باشد نیز به این حالت مربوط می‌شود. دوم آنکه تعداد مشاهدات کافی نباشد. گره‌ای که دارای شرایط شاخه‌بندی نباشد یک برگ نامیده می‌شود. برای تعریف درجه توافق می‌توان پارامترهای زیر را تعریف نمود:

خطای مجاز برای گره (مسئله PR)، واریانس مجاز (مسئله RA) و آستانه برای مشاهدات کیفی [۷].

۴-۳- معیارهای انتخاب صفت

معیارهای مختلفی برای تعیین صفتی که شکاف باید بر اساس آن انجام شود، وجود دارد، مانند:

- بهره اطلاعاتی[1]
- نسبت بهره[2]
- شاخص جینی[3]

در اینجا تنها به معرفی بهره اطلاعاتی می‌پردازیم. بهره اطلاعاتی؛ اطلاعات مورد نیاز برای طبقه‌بندی یک مؤلفه در D برابر:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (۱۰)$$

که در آن p_i احتمال آن است که یک مؤلفه دلخواه در D متعلق به طبقه C_i باشد که این احتمال به صورت $|C_{i,D}|/|D|$ تخمین زده می‌شود. ($|D|$ و $|C_{i,D}|$ تعداد مؤلفه‌ها در D و $C_{i,D}$ را نشان می‌دهد). تعداد طبقه‌های موجود m است.

در واقع (Info D) همان آنتروپی یا بی‌نظمی[4] می‌باشد. فرض می‌شود، صفتٌ A دارای v مقدار متمایز به صورت $\{a_1,a_2,\ldots,a_v\}$ باشد یا به عبارت دیگر A یک صفت گسسته باشد.

اگر بخواهند D را بر حسب صفت A بشکافند، V بخش یا زیرمجموعه مانند $\{D_1,D_2,\ldots,D_v\}$ حاصل می‌شود که در آن D_j شامل مؤلفه‌هایی از D است که مقدار صفت A در آنها برابر a_j خواهد بود. اگر فرض شود که D در گره‌ای چون N قرار داشته

[1] Information Gain
[2] Gain Ratio
[3] Gini Index
[4] entropy

dataacademy.ir

باشد آنگاه این بخش‌ها متناظر با شاخه‌هایی هستند که از N خارج می‌شوند. اطلاعات مورد نیاز برای طبقه‌بندی یک مؤلفه از D بر حسب صفت A برابر است با:

$$Info_A(D) = \sum_{j=1}^{v} |D_j| / |D| * Info(D_j) \quad (۱۱)$$

عبارت $|D_j|/|D|$ در واقع وزن بخش j را نشان می‌دهد. اطلاعات حاصل از انشعاب بر حسب صفت A را به صورت رابطه (۱۲) تعریف می‌نمایند:

$$Gain(A) = Info(D) - Info_A(D) \quad (۱۲)$$

هر چه مقدار بهره صفت A (Gain(A)) بیشتر باشد یا به عبارت دیگر هر چه (InfoA(D)) کمتر باشد صفت A به عنوان صفت شکاف انتخاب می‌شود [۲] و [۸].

۴-۴- پارامترهای کیفیت درخت

فرض شود یک درخت تصمیم و نمونه‌ای از N شئ موجود است. امکان انتخاب دو نوع اصلی از پارامترها توضیح دهنده کیفیت یک درخت وجود دارد. نوع اول پارامترهای دقت هستند و نوع دوم پارامترهای پیچیدگی درخت. پارامترهای دقت یک درخت را می‌توان با کمک نمونه تعریف کرد و کیفیت تقسیم اشیاء در کلاس‌های مختلف (در مورد یک مسئله تشخیص)، یا اندازه بزرگی خطا (در مورد یک مسئله آنالیز رگرسیون) را تعیین نمود. عدد نسبی (فراوانی) خطاها به معنای کسری از اشیاء است که توسط درخت بطور اشتباه به یک کلاس نسبت داده شده:

$$\overline{P_{err}} = \frac{N_{err}}{N} \quad (۱۳)$$

که در آن:

$$N_{err} = \sum_{S=1}^{M} \sum_{\substack{i=1 \\ i \neq \hat{Y}(S)}}^{K} N_i^S \quad (۱۴)$$

که در آن نیز K تعداد کلاسها می‌باشد. واریانس نسبی برای یک درخت تصمیم را می‌توان از فرمول (۱۵) محاسبه نمود:

$$d_{om} = \frac{d_{oc}}{d_o} \quad (۱۵)$$

$$d_{oc} = \frac{1}{N} \sum_{S=1}^{M} \sum_{i \in Data^s} (\hat{Y}(S) - y^i)^2 \quad (۱۶)$$

که در آن d_{oc} واریانس باقیمانده است. واریانس اولیه بصورت:

$$d_o = \frac{1}{N} \sum_{i=1}^{N} (y^i - \bar{y})^2 \quad (۱۷)$$

تعریف می‌شود و $\bar{y}$ به عنوان میانگین می‌باشد.

پارامترهای پیچیدگی، خصوصیات شکل درخت را تعیین می‌کنند و به نمونه بستگی ندارند. برای مثال، پارامترهای پیچیدگی یک درخت به صورت تعداد برگ‌های درخت، تعداد گره‌های داخلی آن و بیشترین طول مسیر از ریشه تا یک برگ می‌باشند. همچنین می‌توان از طول یک مسیر خارجی استفاده نمود که به صورت تعداد شاخه‌هایی تعریف می‌شود که یک درخت کامل را تشکیل می‌دهند.

پارامترهای پیچیدگی و دقت با هم دارای پیوستگی داخلی هستند؛ به عنوان یک قانون می‌توان گفت درختی که پیچیده‌تر باشد دارای دقت بیشتری است (در درختی که هر برگ آن نماینده یک شئ باشد بیشترین میزان دقت وجود دارد).

اگر دیگر شرایط یکسان باشند، درختی که پیچیدگی کمتری داشته باشد، ترجیح داده می‌شود. چنین درختی مدل ساده‌تری از پدیده مورد تحقیق را بدست می‌دهد و تفسیرهای بعدی (توضیح مدل) را آسان می‌کند. علاوه بر این، از تحقیقات تئوری چنین بر می‌آید که در صورت کوچک بودن اندازه نمونه (در مقایسه با تعداد خصوصیات) درخت‌هایی که بیش از حد پیچیده باشند ناپایدار هستند یعنی دارای تعداد خطاهای بیشتری برای مشاهدات جدید خواهند بود. از طرف دیگر، روشن است که یک درخت خیلی ساده نیز امکان رسیدن به پیش‌بینی خوبی را فراهم نمی‌کند. بنابراین، در انتخاب بهترین درخت تصمیم باید به یک "توافق" معینی بین پارامترهای دقت و پیچیدگی برسد.

برای رسیدن به چنین توافقی مثلاً می‌توان از شرط $Q = p + \alpha M$ برای کیفیت استفاده شود. در آن p یک پارامتر دقت و α یک پارامتر معلوم هستند. بهترین درخت با توجه به شرط باید دارای کمترین مقدار Q باشد. از روشی که در آن بیشترین پیچیدگی مجاز برای درخت تعیین می‌شود، بطور همزمان با جستجوی دقیق‌ترین درخت هم می‌توان استفاده نمود [۷].

۵- تخمینی از کیفیت بر روی یک نمونه کنترل

"نمونه کنترل (تست)" به نمونه‌ای گفته می‌شود که برای ساختن یک درخت بکار برده نمی‌شود، بلکه برای تخمین زدن کیفیت یک درخت ساخته شده بکار می‌رود. دو پارامتر محاسبه می‌شود؛ این دو پارامتر تعداد نسبی خطاها برای مسائل تشخیص، و واریانس نمونه کنترل برای مسائل آنالیز رگرسیون هستند. از آنجا که این نمونه در ساخت درخت تصمیم نقشی ندارد، این پارامترها خطای نامعلوم "واقعی" را بهتر نشان می‌دهند. هر چه اندازه نمونه کنترل بزرگتر باشد، درجه تخمین هم بالاتر خواهد بود.

در یک مسئله تشخیص، تحت شرایط مستقل بودن مشاهدات، فراوانی خطاها از توزیع دوجمله‌ای بدست می‌آید.

dataacademy.ir

بنابراین، با دانستن تعداد خطاها در نمونه کنترل می‌توان بازه اطمینانی را پیدا کرد که به احتمال معینی تعداد خطاهای کلاس‌بندی اشتباه به آن بازه تعلق دارد.

۶- ارزیابی درخت ایجاد شده

برای ارزیابی درخت ایجاد شده توسط روش CART یا هر روش دیگری معیارهایی وجود دارند. از مهم‌ترین و اصلی‌ترین این معیارها نرخ خطا در درخت می‌باشد.

رشد درخت بر اساس شاخص جینی از همان گره ریشه، که اولین گره بوده و در برگیرنده تمام مشاهدات است، آغاز شده و برای هر درختی که ایجاد می‌شود، هزینه دسته‌بندی اشتباه آن (که می‌توان از آن به عنوان شاخص خوبی برازش یاد کرد) طبق رابطه (۱۸) محاسبه می‌شود:

$$\text{misclassification cost} = \sum_{t=1}^{T} P(t)\left[1 - \sum_{j=1}^{J} P^{2}(j|t)\right] \quad (۱۸)$$

که در آن P(t)، سهم مشاهدات موجود در گره نهایی t از کل مشاهدات بوده و T، تعداد گره‌های نهایی است. رابطه فوق نمایانگر آن دسته از داده‌هایی است که به اشتباه در دسته‌های غیر مرتبط با خود، دسته‌بندی شده‌اند.

برای محاسبه نرخ خطا در درخت، ابتدا می‌بایست نرخ خطا در هر شاخه بدست آید. نرخ خطا در هر برگ عبارت است از؛ نسبت تعداد رکوردهایی که کلاس یا دسته آنها درست پیش‌بینی نشده است.

جهت برآورد نرخ خطای کل درخت، مجموع وزنی نرخ خطاهای برگ‌ها بدست آورده می‌شود (وزن هر برگ در واقع نسبت جمعیّت آن برگ به کل جمعیّت رکوردها می‌باشد). کیفیّت درخت حاصله نیز مهم خواهد بود.

به منظور جلوگیری از تولید قانون‌های بی‌کیفیّت در بعضی از شاخه‌ها، قطع (هرس) صورت می‌گیرد. این کار با آن‌که نرخ خطا را افزایش می‌دهد ولی از ایجاد بعضی قانون‌های ناکارا جلوگیری می‌نماید. همچنین باید به این نکته توجه داشت که باید قطع کردن به نحوی صورت گیرد که خطا از مقدار معینی بیشتر نشود [۷].

۷- بحث درباره‌ی نتایج

فرض شود تعدادی رکورد وجود دارد که دسته آنها از قبل معلوم می‌باشد (متغیر وابسته در آنها معلوم است). هدف تهیه درختی است که بتوان بوسیله آن متغیر وابسته یا همان کلاس را برای یک رکورد جدید پیش‌بینی و تعیین نمود. روش CART شاخه‌های خود را به صورت دوتایی و تنها بر اساس یک فیلد (متغیر مستقل) ایجاد می‌کند. یعنی هر گروه غیر برگ آن، به دو گروه دیگر تفکیک می‌گردد.

اولین قدم، پاسخ به این سئوال است که کدامـیک از فیلدها بهترین شاخه را تولید می‌کند. بهترین ایجاد شاخه هنگامی رخ می‌دهد که شاخه‌های حاصل طوری باشند که در هر شاخه یک کلاس بر سایر کلاس‌ها غلبه کند.

معیار جهت ارزیابی شاخه‌ها، گوناگونی نام دارد. برای محاسبه گوناگونی در یک مجموعه از رکوردها روش‌های بسیاری وجود دارد که در تمامی آنها گوناگونی زیاد عبارت است مجموعه‌هایی که از کلاس‌های گوناگون در خود داشته باشند و گوناگونی کم عبارت است از مجموعه‌هایی که اعضای یک کلاس در آن بر سایر کلاس‌ها غلبه کند و بهترین نحوه ایجاد شاخه آن است، گوناگونی در مجموعه‌ها را تا حد امکان کم کند.

در مرحله بعد دو شاخه وجود دارد که هرکدام دارای یکـسری رکورد می‌باشند (هر یک از رکوردهای گره بالاتر در یکی از شاخه‌ها قرار گرفته است). حال برای هر شاخه مثل قبل عمل می‌گردد. یعنی برای هر یک از آنها دوباره فیلد طوری انتخاب می‌شود که بتوان بهترین شاخه‌های جدید را با حداقل گوناگونی ایجاد نمود. این مرحله آنقدر ادامه می‌یابد تا در هر زیر شاخه گره‌ای تولید شود که ایجاد شاخه جدید در آن گره مقدار گوناگونی را کاهش قابل توجه‌ای ندهد. به این گره نهایی برگ گفته می‌شود [۹]. برای جداسازی هر گره به دو زیر گره، شاخص‌های مختلفی وجود دارد که معروف‌ترین آن برای داده-های اسمی، شاخص جینی است و به شکل رابطه (۱۹) تعریف می‌شود:

$$P(j|m) = \frac{P(j,m)}{P(m)}, P(j,m) = \frac{\pi(j)N_j(m)}{N_j}, P(m) = \sum_{j=1}^{J} P(j,m) \quad (۱۹)$$

$$\mathbf{Gini(m) = 1 - \sum_{j=1}^{J} P^2(j|m)} \quad (۲۰)$$

که در آن J تعداد دسته‌ها یا همان متغیرهای هدف، π (j)، احتمال اولیه مربوط به دسته j و توسط تصمیم‌گیرنده مشخص می‌شود. $N_j(m)$ تعداد مشاهدات مربوط به دسته j در گره m، N_j تعداد کل مشاهدات مربوط به کلاس j در گره ریشه، P(j|m)، احتمال قرارگیری مشاهدات مربوط به دسته j در گره m و Gini(m)، که همان شاخص جینی است، معرف عدم خلوص یا ناهمگنی در گره m است.

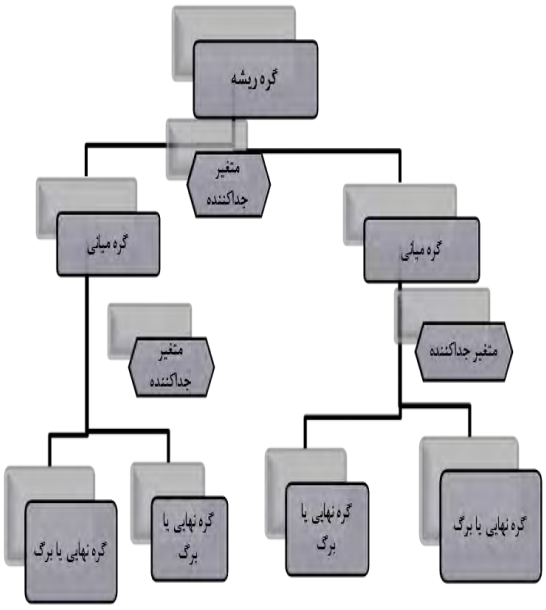
dataacademy.ir

به این معنی که مثلاً اگر همه مشاهدات در یک گره از یک دسته باشند، Gini(m)، برابر صفر و مبین کمترین ناخالصی و به عبارت دیگر بیشترین خلوص در گره است و برعکس، بیشترین مقدار Gini(m)، زمانی حاصل می‌شود که از همه مشاهدات به یک نسبت در گره وجود داشته باشند. شاخص جینی در هر گره برای تمام متغیرها محاسبه شده و متغیری به عنوان متغیر جداکننده انتخاب می‌شود که کمترین مقدار برای جینی از آن بدست آید. احتمال اولیه، مبیّن سهم هر یک از دسته‌ها در جامع مرجع است. رشد درخت بر اساس شاخص جینی از همان گره ریشه، که اولین گره بوده و در برگیرنده تمام مشاهدات است، آغاز شده و برای هر درختی که ایجاد می‌شود، هزینه دسته‌بندی اشتباه آن - که می‌توان از آن به عنوان شاخص خوبی برازش یاد کرد - طبق رابطه (۲۰) محاسبه می‌شود که در آن P(t)، سهم مشاهدات موجود در گره نهایی t از کل مشاهدات بوده و T، تعداد گره‌های نهایی است. رابطه فوق نمایانگر آن دسته از داده-هایی است که به اشتباه در دسته‌های غیر مرتبط با خود، دسته-بندی شده‌اند [۱۰].

برای ارزیابی درخت ایجاد شده توسط روش CART یا هر روش دیگری معیارهایی وجود دارند. از مهمترین و اصلی‌ترین این معیارها نرخ خطا در درخت می‌باشد. برای محاسبه نرخ خطا در درخت ابتدا می‌بایست نرخ خطا در هر شاخه بدست آید. نرخ خطا در هر برگ عبارت است از؛ نسبت تعداد رکوردهایی که کلاس یا دسته آنها درست پیش‌بینی نشده است.

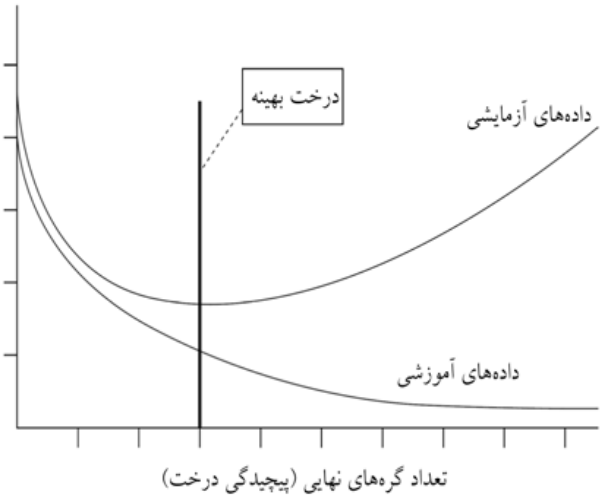
برای محاسبه نرخ خطای کل درخت، مجموع وزنی نرخ خطاهای برگ‌ها بدست آورده می‌شود (وزن هر برگ در واقع نسبت جمعیّت آن برگ به کل جمعیّت رکوردها می‌باشد). کیفیّت درخت حاصله نیز مهم خواهد بود.

جهت جلوگیری از تولید قانون‌های بی‌کیفیّت در بعضی از شاخه‌ها قطع (هرس) صورت می‌گیرد. این کار با آن‌که نرخ خطا را افزایش می‌دهد ولی از ایجاد بعضی قانون‌های ناکارا جلوگیری می‌نماید. همچنین باید به این نکته توجه داشت که باید قطع کردن به نحوی صورت گیرد که خطا از مقدار معینی بیشتر نشود [۹].



شکل (۳): شکل شماتیکی از ساختار کلّی یک درخت تصمیم

درخت بهینه درختی است که کمترین هزینه دسته‌بندی اشتباه را برای داده‌های آزمایشی داشته باشد [۱۰].



شکل (۴): اهمیّت پیچیدگی درخت و هزینه دسته‌بندی اشتباه برای داده‌های آموزشی و آزمایشی

۸- نتیجه‌گیری

کارت که یک مدل ناپارامتری و بدون هرگونه پیش فرض در خصوص رابطه بین متغیرهای مستقل و متغیر هدف است و از روش‌های مهم داده‌کاوی می‌باشد، به طور گسترده در تجارت، صنعت، مهندسی و سایر علوم استفاده می‌شود. الگوریتم طبقه-بندی و درخت رگرسیون سه وظیفه مهم را در بر می‌گیرد؛ اول، این‌که چگونه در هر مرحله داده‌ها را بخش‌بندی نماید. دوم، چه

dataacademy.ir

زمانی بخش‌بندی را متوقف نماید. آخرین وظیفه، چگونگی پیش‌بینی مقدار y برای هر x در یک بخش‌بندی است.

عمل شاخه‌بندی پایه ساخت درخت است. در الگوریتم CRT هر گروه غیر برگ آن، به دو گروه دیگر تفکیک می‌گردد. در این مقاله، قاعده‌ی جینی و قاعده‌ی دوتایی بررسی شده است. گره‌ای که دارای شرایط شاخه‌بندی نباشد یک برگ نامیده می-شود.

برای تعریف درجه توافق (قانون توقف) می‌توان پارامترهای؛ خطای مجاز برای گره (مسئله PR)، واریانس مجاز (مسئله RA) و آستانه برای مشاهدات کیفی را تعریف نمود.

معیارهای؛ بهره اطلاعاتی، نسبت بهره و شاخص جینی برای تعیین صفتی که شکاف باید بر اساس آن انجام شود، وجود دارد. مهمترین و اصلی‌ترین معیارها جهت کنترل کیفیت درخت ایجاد شده، معیار نرخ خطا در درخت می‌باشد.

۹- مراجع

[۱] س. م. شرفی، ح. ر. اسماعیلی، کاربرد داده‌کاوی در پیش‌بینی عیب سطحی فولاد، چهارمین کنفرانس داده‌کاوی ایران، دانشگاه صنعتی شریف، آذر ماه ۱۳۸۹.

[۲] ع. پاک گوهر، بررسی میزان ایمنی کمربند ایمنی و تأثیر آن در کاهش صدمات جسمی و جانی در تصادفات رانندگی، پایان نامه کارشناسی ارشد، دانشکده آمار، دانشگاه آزاد اسلامی واحد مشهد، مشهد، ۱۳۸۵.

[۳] ا. ساکی، ا. حاجی زاده، ن. تهرانیان، بررسی ریسک فاکتورهای سرطان های پستان با استفاده از تحلیل مدل های درختی، *افق دانش؛ فصلنامه ی دانشگاه علوم پزشکی و خدمات بهداشتی درمانی گناباد، صفحات ۶۰ تا ۶۹*، دوره ۱۷، شماره ۱، بهار ۱۳۹۰.

[4] W.-Y. Loh, "Classification and Regression Tree Methods,," *In Encyclopedia of Statistical in Quality and Reliability,* pp. pp. 315- 323, 2008.

[5] A. Dobra, "Classification and Regression Tree Construction," *Thesis Proposal,* p. 0, 2002.

[6] R. Timofeev, "Classification and Regression Trees (CART) Theory and Applications,," *A Master Thesis,Center of Applied Statistics and Economics, Hambodelt University,* p. 0, 2004.

[۷] م .کاظمی، ارزیابی رفتار رانندگان در مواجه با علائم ترافیکی، پایان نامه کارشناسی ارشد، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی واحد زنجان، زنجان، ۱۳۹۰.

[۸] ع. ر. پاک‌گوهر، ع. صادقی‌کیا، تحلیل داده‌های آماری تصادفات رانندگی بوسیله درخت تصمیم، *فصلنامه مطالعات مدیریت ترافیک، سال سوم، شماره۸*، بهار ۱۳۸۷.

[۹] ج. محجوبی، ا. ا. شهیدی، پیش بینی پارامترهای امواج ناشی از باد در بندر امیرآباد به کمک درخت های تصمیم رگرسیونی، *چهارمین کنگره ملّی عمران*، تهران، اردیبهشت ۱۳۸۷.

[۱۰] ا. شریعت مهیمنی، ع. توکلی کاشانی، تعیین شدت مصدومیت ناشی از تصادف‌ها در راه‌های دوخطه برون‌شهری با استفاده از مدل‌های داده‌کاوی، پژوهشنامه حمل و نقل، سال هفتم، شماره۲، تهران، تابستان ۱۳۸۹.