فصلنامه تأمین اجتماعی ۶۰
سال شانزدهم ● شماره سوم ● تابستان ۱۴۰۰

طبقه‌بندی داده‌های نظرسنجی بیمارستان تأمین اجتماعی شهدای کارگر یزد با الگوریتم درخت تصمیم

صمد شهریاری[۱]

چکیده

هدف: داده‌کاوی علم و تکنیک‌هایی است که برای تجزیه‌وتحلیل داده‌ها به‌منظور کشف و استخراج الگوهای ناشناختهٔ قبلی استفاده می‌شود. همچنین، به‌عنوان بخش اصلی فرآیند کشف دانش در پایگاه‌های داده در نظر گرفته می‌شود. هدف اصلی ما ساخت یک مدل طبقه‌بندی کارآمد با دقت بالا برای بهبود کارایی و اثربخشی است.

روش: در این مقاله یک تکنیک یادگیری نظارت‌شده به‌منظور ایجاد یک درخت تصمیم برای داده‌های نظرسنجی بیمارستان تأمین اجتماعی شهدای کارگر یزد معرفی می‌کنیم. هدف اصلی ساخت یک مدل طبقه‌بندی کارآمد با دقت بالا برای بهبود کارایی و اثربخشی فرآیند پذیرش است. برای ساخت درخت تصمیم از الگوریتم CART و بسته rpart موجود در زبان برنامه‌نویسی R استفاده شده است و مدل نهایی با استفاده از روش‌های رایج ارزیابی، ارزیابی شد.

نتیجه: طبق نتیجهٔ به دست آمده، مهم‌ترین طبقه (از سمت راست) طبقهٔ اول است؛ زیرا حدود ۸۴ درصد از داده‌ها را شامل می‌شود. این طبقه نشان می‌دهد که اگر میزان رضایت از پزشک بزرگتر یا مساوری ۳ و همچنین، میزان رضایت از کارکنان پذیرش بزرگتر یا مساوی ۴ باشد، ۸۴ درصد از مراجعه‌کنندگان در صورت نیاز، مجدداً به بیمارستان مراجعه می‌کنند.

واژگان کلیدی: ارزیابی مدل، داده‌کاوی، درخت تصمیم، طبقه‌بندی

۱. سازمان تأمین اجتماعی، دانشجوی کارشناسی ارشد علم داده‌ها، دانشگاه علامه طباطبایی smd.shahriari@yahoo.com

۱. مقدمه

داده‌کاوی ترجمه عبارت لاتین Data Mining[1] و به معنای تحت‌اللفظی «کاویدن داده» است. کلمه Mining[2] در معنای تحت‌اللفظی خود، یعنی «استخراج از معدن» به کار می‌رود. در واقع، واژهٔ «داده‌کاوی» نشان می‌دهد که حجم انبوه اطلاعات مانند یک معدن عمل می‌کند و از ظاهر آن مشخص نیست چه عناصر گرانبهایی در عمق آن وجود دارد. تنها با کندوکاو و استخراج این معدن است که می‌توان به آن عناصر گرانبها دست پیدا کرد.

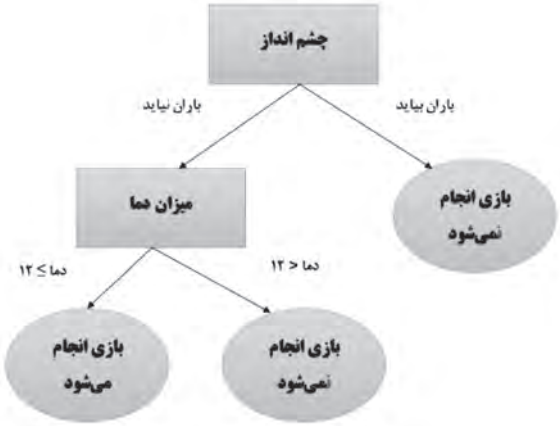
طبقه‌بندی یک روش داده‌کاوی است که موارد موجود در یک مجموعه را به دسته‌ها یا کلاس‌های هدف اختصاص می‌دهد. برای این منظور تکنیک‌های مختلف طبقه‌بندی (درخت تصمیم، نزدیک‌ترین همسایه‌ها، ماشین بردار پشتیبان و...) استفاده می‌شود.

در داده‌کاوی، درخت تصمیم (که به آن درخت طبقه‌بندی نیز گفته می‌شود) یک مدل پیش‌بینی است که می‌تواند برای مدل طبقه‌بندی استفاده شود. درختان طبقه‌بندی به‌عنوان یک تکنیک اکتشافی مفید هستند و در بسیاری از زمینه‌ها مانند مالی، بازاریابی، پزشکی و مهندسی استفاده می‌شوند.

استفاده از درخت تصمیم به‌دلیل سادگی و شفافیت در داده‌کاوی بسیار محبوب است. درخت‌های تصمیم معمولاً به‌صورت گرافیکی به عنوان یک ساختار سلسله مراتبی نشان داده می‌شوند که تفسیر آنها را نسبت به سایر تکنیک‌ها آسان‌تر می‌کند. این ساختار عمدتاً شامل یک گره شروع (به نام ریشه) و گروهی از شاخه‌ها (شرایط) است که به گره‌های دیگر منتهی می‌شود تا زمانی که به گره برگ که حاوی تصمیم نهایی این مسیر است برسیم و در نهایت، هر سرنخ یک طبقه‌بندی را اختصاص می‌دهد. درخت تصمیم یک مدل خود توضیحی است؛ زیرا نمایش آن بسیار ساده است.

شکل ۱ مثالی برای یک درخت تصمیم‌گیری ساده برای طبقه‌بندی «بازی تنیس» رل نشان می‌دهد. با توجه به این شکل، به‌سادگی می‌توان تصمیم گرفت که بازی تنیس انجام شود یا خیر که این تصمیم‌گیری بر اساس دو ویژگی چشم‌انداز، میزان دما و یک متغیر هدف که مقادیرش بله یا خیر هستند، صورت می‌پذیرد.

۱. داده‌کاوی
۲. استخراج‌کردن



شکل ۱. مثال درخت تصمیم

طبق شکل ۱، اگر چشم‌انداز طوری باشد که باران نیاید و میزان دما کمتر از ۱۲ درجه باشد، تصمیم می‌گیریم که تنیس بازی نکنیم؛ زیرا مسیری که از گره ریشه شروع می‌شود با کلاس بازی انجام نمی‌شود به پایان می‌رسد.

در این مقاله، ما یک تکنیک یادگیری نظارت‌شده برای ساخت یک مدل درخت تصمیم برای داده‌های نظرسنجی بیمارستان تأمین اجتماعی شهدای کارگر یزد را معرفی می‌کنیم تا ابزاری برای بهبود کارایی و اثربخشی فرآیند پذیرش ارائه کنیم. تجزیه‌وتحلیل این سوابق برای تعریف رابطه بین داده‌های بیماران و وضعیت مراجعه مجدد مورد نیاز است.

این مقاله در پنج بخش تنظیم شده است. در بخش ۲، مدل درخت تصمیم ارائه شده است. بخش ۳، جزئیات مختصری در مورد روش‌های رایج مورد استفاده برای ارزیابی مدل طبقه‌بندی ارائه می‌دهد. در بخش ۴، نتایج تجربی با توجه به نتایج مدل و دیدگاه سیستم پذیرش ارائه و تحلیل می‌شوند. در نهایت، نتیجه‌گیری این کار در بخش ۵ ارائه شده است.

۲. مدل درخت تصمیم

درخت تصمیم طبقه‌بندی‌کنند ای است که به‌صورت یک پارتیشن بازگشتی از فضای ورودی بر اساس مقادیر متغیرها بیان می‌شود. همان‌طور که قبلاً گفته شد، هر گره داخلی فضای نمونه را به دو یا چند فضای فرعی با توجه به عملکرد مشخصی از مقادیر متغیر ورودی تقسیم می‌کند. هر برگ به یک کلاس اختصاص داده می‌شود که مناسب‌ترین یا پرتکرارترین مقدار هدف را نشان می‌دهد.

نمونه‌ها با پیمایش درخت از گره ریشه تا یک برگ بر اساس نتیجهٔ گره‌های آزمایشی در طول این مسیر

طبقه‌بندی می‌شوند. هر مسیر را می‌توان با پیوستن به تست‌های این مسیر به یک قانون تبدیل کرد. برای مثال، یکی از مسیرهای شکل ۱ را می‌توان به قانون تبدیل کرد: «اگر چشم‌انداز طوری باشد که باران نیاید و دما بالاتر از ۱۲ درجه باشد، می‌توانیم تنیس بازی کنیم». قوانین به‌دست‌آمده برای توضیح یا درک سیستم به‌خوبی استفاده می‌شود.

الگوریتم‌های زیادی برای یادگیری درخت تصمیم از یک مجموعه داده ارائه شده است؛ اما ما از الگوریتم CART[1] به‌دلیل سادگی آن برای پیاده‌سازی استفاده خواهیم کرد. در این بخش، از الگوریتم CART برای ساخت درخت تصمیم و برخی از توابع متداول مورد استفاده برای تقسیم فضای ورودی بحث می‌کنیم.

۲-۱. الگوریتم CART

الگوریتم CART نیازی به فرض توزیع داده ندارد و اجازه می‌دهد تا ورودی‌ها مخلوطی از متغیرهای طبقه‌ای و پیوسته باشند. این برای مجموعهٔ داده‌های بزرگ مفید است؛ زیرا نتایج ارزشمندی را با چند متغیر مهم ارائه می‌دهد.

الگوریتم CART از طریق فرآیند زیر کار می‌کند:

- بهترین نقطهٔ تقسیم هر متغیر را پیدا می‌کند.
- بر اساس بهترین نقاط تقسیم هر متغیر در مرحلهٔ ۱، «بهترین» نقطهٔ تقسیم جدید شناسایی می‌شود.
- متغیر انتخاب‌شده را بر اساس «بهترین» نقطهٔ تقسیم، تقسیم‌بندی می‌کند.
- تا زمانی که یک قانون توقف برآورده شود یا تقسیم مطلوب دیگری در دسترس نباشد، تقسیم‌بندی را ادامه می‌دهد.

۳. ارزیابی مدل

یک متغیر دودویی را در نظر بگیرید (یعنی فقط دو مقدار دارد: مثبت و منفی). داده‌های خروجی یک مدل طبقه‌بندی تعداد نمونه‌های صحیح و نادرست با توجه به کلاس شناخته‌شدهٔ قبلی آنها است. این تعداد در ماتریس درهم‌ریختگی، همان‌طور که در جدول ۱ نشان داده شده، رسم شده است.

جدول ۱. ماتریس درهم ریختگی (متغیر دودویی)

		برچسب پیش‌بینی شده	
		مثبت	منفی
برچسب شناخته شده	مثبت	FN	FN
	منفی	FP	TN

1. Classification and Regression Trees

همان‌طور که در جدول ۱ مشخص است، TP[1] نشان می‌دهد که ما پیش‌بینی کرده‌ایم که مثبت باشد و مثبت بوده است. FP[2] نشان می‌دهد که ما پیش‌بینی کرده‌ایم مثبت باشد، ولی منفی بوده است. TN[3] نشان می‌دهد که پیش‌بینی کرده‌ایم منفی باشد و منفی بوده است. FN[4] نشان می‌دهد که پیش‌بینی کرده‌ایم منفی باشد، ولی مثبت بوده است.

معیارهای ارزیابی زیادی برای ارزیابی عملکرد طبقه‌بندی‌کننده بر اساس ماتریس درهم‌ریختگی حاصل از آزمایش استفاده می‌شود. ما با جزئیات بیشتر برخی از اقدامات رایج مورداستفاده را که بعداً در آزمایش خود مورد استفاده قرار می‌گیرند، شرح خواهیم داد. پرکاربردترین معیاری که برای بررسی دقت طبقه‌بندی استفاده می‌شود، معیار Accuracy[5] است که اثربخشی یک طبقه‌بندی‌کننده را بر اساس درصد نمونه‌های به‌درستی پیش‌بینی‌شدهٔ آن مانند (۱) ارزیابی می‌کند.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (۱$$

Recall[6] (R), Precision[7] (P) مقادیر دیگری هستند که براساس ماتریس درهم‌ریختگی برای ارزیابی مدل استفاده می‌شوند. Recall نسبت مقادیری است که کلاس مثبت داشته و مثبت پیش‌بینی شده‌اند. از طرف دیگر، Precision احتمالی است که یک پیش‌بینی مثبت به‌درستی پیش‌بینی شده باشد و در (۲) نشان داده شده است.

$$R = \frac{TP}{TP+FN} \quad , \quad P = \frac{TP}{TP+FP} \quad (۲$$

Precision , Recall می‌توانند با همدیگر ترکیب شوند و مقدار جدیدی بسازند که F1measure نامیده می‌شود و در (۳) نشان داده شده است.

$$F1measure = \frac{2 * R * P}{R+P} \quad (۳$$

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

1. True Possitive
2. False Possitive
3. True Negative
4. False Negative

۵. دقت
۶. پوشش
۷. درستی

۴. آزمایش

۴-۱. مجموعه داده

در این مقاله از مجموعه دادهٔ نظرسنجی سرپایی سال ۱۴۰۰ بیمارستان تأمین اجتماعی شهدای کارگر یزد که از سامانه CRM[1] استخراج شده است، استفاده کرده‌ایم. این مجموعه داده شامل پنج متغیر و هر متغیر دارای ۴۷۰ مشاهده است که متغیر کلاس که همان MM است را با دو مقدار ۱ (مراجعهٔ مجدد به بیمارستان) یا صفر (عدم مراجعه مجدد به بیمارستان) نشان می‌دهد. متغیر کلاس ۹۲ درصد به‌عنوان مراجعه مجدد و ۸ درصد به‌عنوان عدم مراجعهٔ مجدد به بیمارستان توزیع می‌شود. جدول ۲، اطلاعات دقیقی در مورد متغیرهای مجموعه داده را نشان می‌دهد.

مجموعه داده به دو بخش اصلی تقسیم می‌شود: مجموعه دادهٔ آموزشی که شامل۳۷۶ مشاهده (حدود ۸۰ درصد) را در خود جای می‌دهد و مجموعه دادهٔ آزمایشی که شامل۹۴ مشاهده (حدود ۲۰ درصد) در خود جای می‌دهد. طبقه‌بندی‌کنندهٔ درخت تصمیم با استفاده از یک مجموعه دادهٔ آموزشی آموخته می‌شود و عملکرد آن روی مجموعه داده‌های آزمایشی دیده‌نشده اندازه‌گیری می‌شود.

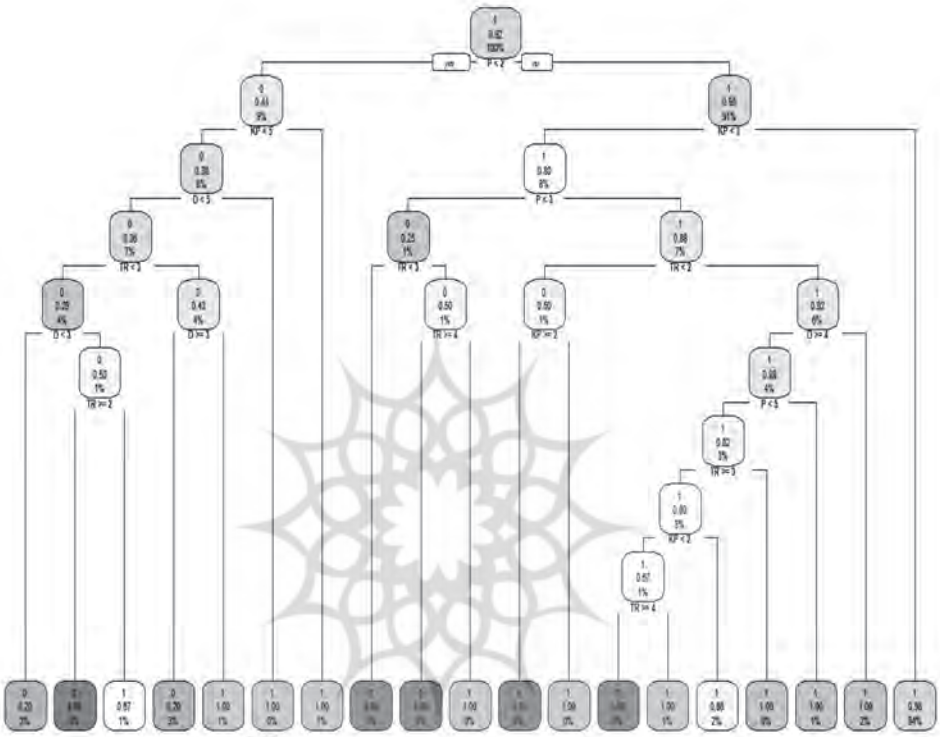
جدول ۲. خلاصه‌ای از متغیرهای مجموعه داده

ویژگی	مقادیر ممکن
MM	آیا در صورت نیاز دوباره به این بیمارستان مراجعه می کنید؟ • ۱ • ۰
TR	میزان رضایت از تسهیلات رفاهی • ۱ • ۲ • ۳ • ۴ • ۵
KP	میزان رضایت از برخورد و خدمات کارکنان پذیرش • ۱ • ۲ • ۳ • ۴ • ۵
P	میزان رضایت از برخورد و خدمات پزشکان • ۱ • ۲ • ۳ • ۴ • ۵
D	میزان رضایت از برخورد و خدمات داروخانه • ۱ • ۲ • ۳ • ۴ • ۵

1- Citizen Relation Management

۴-۲. رسم مدل درخت تصمیم روی داده‌ها

برای رسم درخت تصمیم از بسته[1] rpart موجود در نرم‌افزار R کمک گرفته‌ایم و نتیجهٔ حاصل‌شده در شکل ۲، قابل‌مشاهده است:



شکل ۲. نمودار طبقه‌بندی درخت تصمیم حاصل از داده‌های نظرسنجی بیمارستان

۴-۳. قوانین استنتاج‌شده از درخت تصمیم

با توجه به شکل ۲، اگر از سمت پایین به بالا حرکت کنیم، هرکدام از شاخه‌ها یک طبقه را نشان می‌دهد. همچنین، در پایین‌ترین سطح نمودار رنگ سبز به مفهوم این است که این طبقه در کلاس ۱ (مراجعهٔ مجدد) و رنگ آبی به مفهوم این است که این طبقه در کلاس صفر (عدم مراجعهٔ مجدد) قرار می‌گیرد. شاخه‌ای (طبقه‌ای) که برای ما حائز اهمیت است، از سمت راست، شاخه اول است که ۸۴ درصد از داده‌ها را در خود جای داده است. این شاخه نشان می‌دهد که اگر $P \geq 3$ و $KP \geq 4$ آنگاه ۸۴ درصد از مراجعه‌کنندگان در صورت نیاز، مجدد به بیمارستان مراجعه می‌کنند.

1. Recursive Partitioning and Regression Trees

جدول ۳. برخی مجموعه قوانین درخت تصمیم

اگر $P < 2$ و $KP < 5$ و $D < 5$ و $TR < 3$ و $D < 3$ آنگاه عدم مراجعه مجدد	٪۳
اگر $P < 2$ و $KP < 5$ و $D < 5$ و $TR \geq 3$ و $D < 3$ آنگاه عدم مراجعه مجدد	٪۳
اگر $P < 2$ و $KP < 5$ و $D < 5$ و $TR \geq 3$ و $D \geq 3$ آنگاه مراجعه مجدد	٪۱

۴-۴. نتایج مدل درخت تصمیم

مقادیر ماتریس درهم‌ریختگی در جدول ۴ نشان داده شده است. مقادیر ماتریس درهم ریختگی با اعمال یک درخت تصمیم روی مجموعه داده‌های آزمایشی تولید می شوند.

جدول ٤. ماتریس درهم‌ریختگی داده‌های آزمایشی

		برچسب پیش‌بینی‌شده	
		مثبت	منفی
برچسب شناخته شده	مثبت	۸۳	۲
	منفی	۴	۵

جدول ۵. مقادیر ارزیابی مدل

مقادیر سنجش
$Accuracy = \frac{35 + 2}{35 + 2 + 2 + 1} = 0.936$
$Recall = \frac{35}{35 + 2} = 0.97$ $Precision = \frac{35}{35 + 1} = 0.95$
$F1measure = \frac{2 * 0.94 * 0.97}{0.94 + 0.97} = 0.95$

مقادیر ارزیابی‌ای که در جدول ۵ نشان داده شده است، همگی مقادیر بالایی دارند و این نشان می‌دهد که مدل طبقه‌بندی‌کنندهٔ ما به‌خوبی عمل کرده است.

۵. نتیجه

در این مقاله تأثیر طبقه‌بندی مدل درخت تصمیم برای داده‌های نظرسنجی بیمارستان تأمین اجتماعی شهدای کارگر یزد را نشان دادیم. نتایج آزمایش نشان می‌دهد که باید بیشتر تمرکز را روی شرایط شاخهٔ اول بگذاریم؛ زیرا این شاخه در مجموع ۸۴ درصد از داده‌ها را در خود جای می‌دهند. همچنین، یک مجموعه قوانین جدید توسط مدل درخت تصمیم به دست آمد که امیدواریم در آینده به پیشرفت بیمارستان شهدای کارگر کمک کند.

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

منابع

- «داده‌کاوی و روش‌ها به زبان ساده».

Available at: https://afaghhosting.net/blog/%D8%AF%D8%A7%D8%AF%D987%-%DA%A9%D8%A7%D9%88%DB8%C-data-mining-%D988%-%D8%B1%D9%88%D8%B4-%D9%87%D8%A7-%D8%A8%D987%-%D8%B2%D8%A8%D8%A7%D986%-%D8%B3%D8%A7%D8%AF%D987%/

- عباس‌نیا، لادن (۱۳۹۹) «درخت تصمیم (Decision Tree) چیست؟-۳ الگوریتم پرکاربرد آن».

Available at: https://amarpishro.com/data-analysis/decision-tree/

- Alharan, A.; Radhwan Alsagheer and A. Al-Haboobi (2017) "Popular Decision Tree Algorithms of Data Mining Techniques: A Review", International Journal of Computer Science and Mobile Computing, 6(6):133-142.
- Chiu. D (2015) Machine Learning with R Cookbook. Birmingham: Packt Publishing.
- Gupta, B.; Aditya Rawat; Akshay Jain; Arpit Arora; Naresh Dhami (2017) "Analysis of Various Decision Tree Algorithms for Classification in Data Mining", International Journal of Computer Applications (0975 – 8887). 163(8):15-19.
- Jijo, B. T.; Adnan Mohsin Abdulazeez (2021) "Classification Based on Decision Tree Algorithms for Machine Learning", Journal of Applied Science and Technology Trends, 2(1):20-28.
- Kotsiantis, Sotiris B. (2007) "Supervised Machine Learning: A Review of Classification Techniques", Emerging Artificial Intelligence Applications in Computer Engineering, 160(1):3-24.
- Mashat, A. F.; Mohammad M. Fouad; Philip S. Yu; Tarek F. Gharib (2012) "A Decision Tree Classification Model for University Admission System", International Journal of Advanced Computer Science and Applications. 3(10):17-21.
- Sharma, H.; Sunil Kumar (2016) "A Survey on Decision Tree Algorithms of Classification in Data Mining", International Journal of Science and Research. 5(4): 2094-2097.
- Sohn, S. Y.; Ji Won Kim (2012) "Decision Tree Based Technology Credit Scoring for Start-up Firms: Korean Case", All Science Journal Classification. 39(4):4007-4012.
- Sulaiman, Maryam A. (2020) "Evaluating Data Mining Classification Methods Performance in Internet of Things Applications", Journal of Soft Computing and Data Mining, 1(2):11-25.