

مروری بر درختان تصمیم فازی و غیر فازی

زهره بهدانی* و وحیده احراری

به یاد روانشاد دکتر بهرام صادقپور گیلده

گروه ریاضی، دانشگاه صنعتی خاتم الانبیاء بهبهان، بهبهان، ایران

دانشگاه فردوسی مشهد، دانشکده علوم ریاضی، گروه آمار

تاریخ پذیرش: ۱۴۰۱/۳/۳۱

تاریخ دریافت: ۱۴۰۰/۱۲/۱

نوع مقاله: علمی-مروری

چکیده

علم داده‌کاوی با گسترش سیستم‌های پایگاه داده و حجم بالای داده‌های ذخیره شده در این سیستم‌ها مطرح شده است، تا بتوان الگوهای مفید در داده‌ها را شناسایی کرده و با در اختیار قرار دادن اطلاعات به کاربران، آنها را در اتخاذ تصمیمات مهم کمک نمود. در علم داده‌کاوی نظریه مجموعه‌های فازی نقش مهمی دارد و باعث پیدایش «داده‌کاوی فازی» شده است. پژوهش‌های متعددی در حوزه داده‌کاوی فازی انجام شده که در این مقاله، نقش مبحث فازی در درختان تصمیم مورد مطالعه قرار گرفته است.

۱ مقدمه

دنیای مدرن کنونی در حقیقت دنیای داده‌گرا می‌باشد و بشر در محاصره داده‌های عددی و غیر عددی می‌باشد. برای اینکه این داده‌ها را بتوان به عنوان ابزار مناسب در زمینه‌های مختلف مورد استفاده قرار داد، نیاز است که آنها را تحلیل و پردازش نمود. در حقیقت فرآیند به کارگیری روش‌شناسی مبتنی بر کامپیوتر برای دریافت دانش و اطلاعات از

عبارات و کلمات کلیدی: مجموعه‌های فازی، داده‌کاوی، درخت تصمیم فازی، تابع عضویت.

Email(s): behdani@bkatu.ac.ir and v.ahrari84@gmail.com.

داده‌ها را داده‌کاوی^۱ گویند. می‌توان داده‌کاوی را تعامل همکاری بین انسان و کامپیوتر دانست که به یک فرآیند جهانی که از چندین مرحله تشکیل شده (پیش پردازش اطلاعات، مصورسازی، یادگیری، تحلیل، انتخاب، ...) اشاره دارد. از مهمترین علوم کاربردی در داده‌کاوی، علم آمار می‌باشد. آمار، اساس اغلب تکنولوژی‌هایی می‌باشد که داده‌کاوی بر روی آن‌ها بنا می‌شود. در داده‌کاوی دو هدف مهم پیش‌بینی و توصیف مدنظر است که هر یک از این اهداف با استفاده از روش‌ها و تکنیک‌هایی محقق می‌شوند. درخت تصمیم^۲، از تکنیک‌های داده‌کاوی است که در بسیاری از برنامه‌های کاربردی برای طبقه‌بندی مسائل به کار گرفته می‌شوند.

درختان تصمیم از عمومی‌ترین روش‌های یادگیری با ناظر می‌باشند. ولی چنانچه داده‌ها دارای کاستی‌ها و مشکلاتی از قبیل اغتشاش، حجم کم نمونه، دقت کم، ارزیابی شخصی و ... باشند، درخت تصمیم کارایی کافی را نخواهد داشت. علاوه بر این، مشکلات دیگری مانند وجود ویژگی‌های عددی پیوسته یا گسسته با اندازه (حجم نمونه) زیاد نیز، عملکرد این درختان را تحت تاثیر قرار خواهند داد. در مواردی که درخت تصمیم دچار نقص می‌شود یک رهیافت جایگزین، ترکیب منطق فازی با درختان تصمیم است. که حاصل آن درختان تصمیم فازی است. لازم به ذکر است که بر خلاف داده‌کاوی کلاسیک، در حال حاضر در داده‌کاوی فازی معیارهای متشکل از مجموعه داده‌های فازی برای مقایسه الگوریتم‌ها وجود ندارد.

مقاله حاضر به بررسی مفهوم درختان تصمیم و منطق فازی می‌پردازد و سپس ترکیب این دو مفهوم یعنی درخت تصمیم فازی را تعریف خواهد کرد و به کاربردها و اهمیت آن خواهد پرداخت. این مقاله به مشارکت بالقوه‌ی نظریه مجموعه‌های فازی و درختان تصمیم پرداخته و حوزه‌های مربوطه را برجسته می‌کند. در این رابطه برخی از مزایای روش‌های فازی برای نمایش و استخراج الگوهای مبهم داده‌کاوی در داده‌ها به طور خاص مورد تاکید قرار می‌گیرند، سپس کاربردهای مختلف از مجموعه‌های فازی در درختان تصمیم ارائه شده است. ابتدا در بخش ۲ به تعریف منطق فازی که در سال ۱۹۶۵ توسط لطفی زاده معرفی شده است خواهیم پرداخت. در بخش ۳ مفهوم داده‌کاوی و درختان تصمیم در داده‌کاوی را تعریف خواهیم کرد و در بخش ۴ درختان تصمیم فازی را مرور خواهیم کرد که

¹Data Mining

²Decision tree

شامل مفاهیم داده‌کاوی فازی و الگوهای فازی کاوی در داده‌ها و استخراج ویژگی فازی و نمایش الگو می‌باشد. برای درک بهتر از مفاهیم مطرح شده، در بخش ۵ یک مثال عددی ذکر کرده و درخت تصمیم فازی متناسب با آن رسم شده است. در پایان، کاربردهایی از درختان تصمیم فازی در بخش ۶، ارائه می‌شود.

۲ منطق فازی

واژه فازی به معنای نادقیق، نرم، مبهم و گنگ است. تئوری فازی یک چارچوب جدید است که توانایی مدل کردن واقعیت را، آنچنان که هست دارد. تئوری فازی توسط پرفسور لطفی عسگرزاده یا زاده در سال ۱۹۶۵ معرفی شد [۳۲] و در حال حاضر کاربردهای فراوانی دارد. زاده به ناتوانی ریاضیات کلاسیک برای پرداختن به مسائل نادقیق دنیای واقعی اشاره کرد و چارچوب جدیدی بنام تئوری فازی را پایه ریزی کرد و مبانی آن را معرفی کرد. هدف اولیه او در آن زمان، توسعه مدلی کارآمدتر برای توصیف فرآیند پردازش زبان‌های طبیعی بود. او مفاهیم و اصطلاحاتی هم‌چون مجموعه‌های فازی، رویدادهای فازی، اعداد فازی و فازی سازی را وارد علوم ریاضیات و مهندسی نمود. اساساً منطق فازی یک منطق چند منظوره است. منطق فازی راهکاری است که با استفاده از آن می‌توان سیستم‌های پیچیده را به آسانی مدل‌سازی کرد. دنیایی که در آن زندگی می‌کنیم، دنیای مبهمات و عدم قطعیت است. درحقیقت منطق فازی، جهان را آن‌طور که هست به تصویر می‌کشد. از ویژگی‌های منطق فازی به موارد زیر می‌توان اشاره کرد:

۱. استدلال‌های دقیق در منطق فازی، به عنوان موارد مرزی استدلال‌های تقریبی تلقی می‌شوند.

۲. در منطق فازی، برای هر پیشامد می‌توان درجه‌ای تعیین کرد یا نسبت داد.

۳. هر سیستم منطقی را می‌توان فازی نمود.

۴. دانش در منطق فازی، مجموعه‌ای از محدودیت‌های تغییر پذیر است.

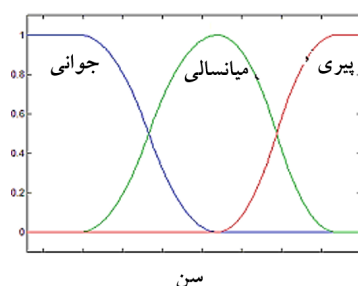
۵. استنتاج، به عنوان فرآیند گسترش محدودیت‌های تغییر پذیر در نظر گرفته می‌شود.

۱.۲ مجموعه‌های فازی

مجموعه‌های فازی در حقیقت شالوده و مبنای منطق فازی است. این نظریه تعمیمی از نظریه کلاسیک مجموعه‌ها در علم ریاضیات است. یک عنصر در تئوری کلاسیک مجموعه‌ها، عضو آن مجموعه است و یا عضو آن نیست و عضویت اعضاء مجموعه از یک الگوی صفر و یک پیروی می‌کند. اما تئوری مجموعه‌های فازی این مفهوم را توسعه می‌دهد و عضویت درجه‌بندی شده را مطرح می‌کند. به این ترتیب که یک عنصر می‌تواند تا درجاتی بین صفر و یک عضو یک مجموعه باشد.

اگر Ω یک مجموعه مرجع باشد، زیر مجموعه فازی A ($\tilde{A}$) از Ω با تابع عضویتش به صورت $\mu_{\tilde{A}}(x)$ نمایش داده می‌شود که، به ازای هر x در Ω مقادیری در بازه $[0, 1]$ تولید می‌کند. در حقیقت تابع عضویت، درجه عضویت نمونه را در مجموعه فازی مربوط به آن بیان می‌کند. اگر $\mu_{\tilde{A}}(x) = 0$ ، آن عضو کاملاً از مجموعه خارج و به عبارتی x متعلق به $\tilde{A}$ نیست و اگر $\mu_{\tilde{A}}(x) = 1$ باشد، آن عضو کاملاً در مجموعه قرار دارد. اگر $\mu_{\tilde{A}}(x) = k$ که $0 < k < 1$ ، می‌گوییم درجه عضویت x در مجموعه فازی $\tilde{A}$ برابر با k است. هرگاه $\mu_{\tilde{A}}(x)$ همواره برابر با ۰ یا ۱ باشد، یک مجموعه دقیق (غیر فازی) از Ω بدست می‌آید. می‌توان نتیجه گرفت مجموعه‌های فازی تعمیمی برای مجموعه‌های کلاسیک هستند بطوریکه هر مجموعه کلاسیک یک مجموعه فازی محسوب می‌شود ولی عکس آن لزوماً برقرار نیست.

یک مجموعه فازی تهی است اگر و فقط اگر، تابع عضویت آن دقیقاً صفر باشد. دو مجموعه فازی $\tilde{A}$ و $\tilde{B}$ مساوی می‌باشند اگر و فقط اگر، برای هر x در Ω داشته باشیم $\mu_{\tilde{A}}(x) = \mu_{\tilde{B}}(x)$. مکمل مجموعه فازی $\tilde{A}$ نیز با $\tilde{A}'$ نشان داده می‌شود که $\mu_{\tilde{A}'}(x) = 1 - \mu_{\tilde{A}}(x)$ روش‌های مختلفی برای تولید توابع عضویت روی مجموعه داده‌ها ارائه می‌شود. از لحاظ مفهومی در ضمن می‌تواند هر مجموعه فازی به صورت تداخلی با درجه‌ای در مجموعه فازی دیگر قرار گیرد. مثلاً در متغیر زبانی سن، صفت جوانی را مد نظر بگیریم، حال با توجه به انتخاب تابع عضویت مانند گوسین، صفت میان سالی با درجه عضویت کم می‌تواند در مجموعه صفت جوانی قرار گیرد و صفت پیری نیز با درجه عضویت کمتری در مجموعه صفت جوانی ظاهر می‌شود (شکل ۱ می‌تواند برای درک این موضوع کمک کند).



شکل ۱: مجموعه فازی جوانی، میانسالی و پیری

۲.۲ اعداد فازی

اعداد فازی تعمیمی از اعداد حقیقی می‌باشند که غالباً برای بیان عدم حتمیت بکار می‌روند. به اینصورت که یک مقدار واحد را نشان نمی‌دهند، بلکه مجموعه‌ای پیوسته از مقادیر ممکن را نشان می‌دهد که هر مقدار وزنی بین صفر و یک دارد که درجه عضویت نام دارد.

تعریف ۱.۲. به مجموعه ی فازی $\tilde{A}$ با تابع عضویت $\mu_{\tilde{A}}(x)$ عدد فازی گفته می‌شود هر گاه دارای خواص زیر باشد:

۱. مجموعه فازی $\tilde{A}$ محدب باشد.

۲. مجموعه فازی $\tilde{A}$ نرمال باشد (یعنی حداقل یک نقطه با درجه عضویت یک داشته باشد).

۳. مجموعه فازی $\tilde{A}$ پیوسته یا پیوسته قطعه وار باشد.

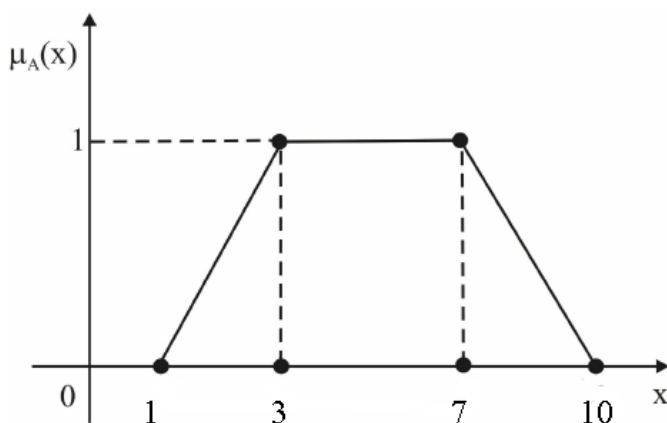
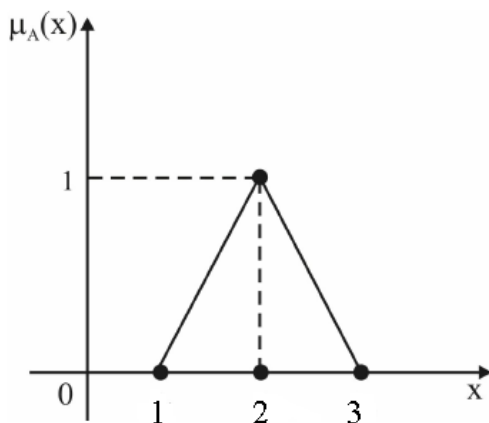
براساس مجموعه مقادارها و شکل تابع عضویت، اعداد فازی به چند دسته تقسیم می‌شوند.

تعریف ۲.۲. یک عدد فازی مثلثی مانند $\tilde{N}$ با سه عدد حقیقی $a < b < c$ تعریف می‌شود. قاعده مثلث بازه $[a, c]$ و راس آن در نقطه $x = b$ است. عدد فازی مثلثی را به صورت $\tilde{N} = (a, b, c)$ نمایش می‌دهیم.

تعریف ۳.۲. یک عدد فازی دوزنقه‌ای مانند $\tilde{M}$ با چهار عدد حقیقی $a < b < c < d$

تعریف می‌گردد. که قاعده دوزنقه بازه $[a, d]$ و راس آن در بازه $[b, c]$ درجه عضویت یک دارد. نمایش این عدد فازی نیز به صورت $\tilde{M} = (a, b, c, d)$ است.

شکل ۳ نمودار بالا عدد فازی مثلثی $\tilde{N} = (1, 2, 3)$ و نمودار پایین عدد فازی دوزنقه‌ای $\tilde{M} = (1, 3, 7, 10)$ را نشان می‌دهد. عدد فازی گوسین یا زنگوله‌ای شکل و



شکل ۲: عدد فازی مثلثی $\tilde{N} = (1, 2, 3)$ (شکل بالا) و عدد فازی دوزنقه‌ای $\tilde{M} = (1, 3, 7, 10)$ (شکل پایین)

اعداد فازی پی نیز از دیگر اعداد فازی هستند که مطالب بیشتر در مورد این داده ها را

می‌توان در [۴] مشاهده کرد.

برای تعمیم عملگرهای حسابی معمولی به عملگرهای حسابی فازی، نیاز به بیان و تشریح یک اصل اساسی است که به اصل توسیع موسوم است. در ادامه این اصل ارائه می‌شود.

تعریف ۴.۲ (اصل توسیع). فرض کنید $X_1, \dots, X_n$ ، مجموعه مرجع و $X = X_1 \times \dots \times X_n$ حاصل ضرب دکارتی آن‌ها باشد. همچنین $A_1, \dots, A_n$ مجموعه فازی به ترتیب از $X_1, \dots, X_n$ باشند. به علاوه $y = f(x_1, \dots, x_n)$ یک نگاشت از X به Y باشد. حاصل عمل f بر n مجموعه فازی $A_1, \dots, A_n$ به صورت مجموعه فازی B از Y با تابع عضویت زیر تعریف می‌شود:

$$B(y) = f(A_1, \dots, A_n)(y) = \begin{cases} \sup_{x_1, \dots, x_n} \min\{A_1(x_1), \dots, A_n(x_n)\} & ; f^{-1}(y) \neq \emptyset, \\ 0 & ; f^{-1}(y) = \emptyset, \end{cases} \quad (1)$$

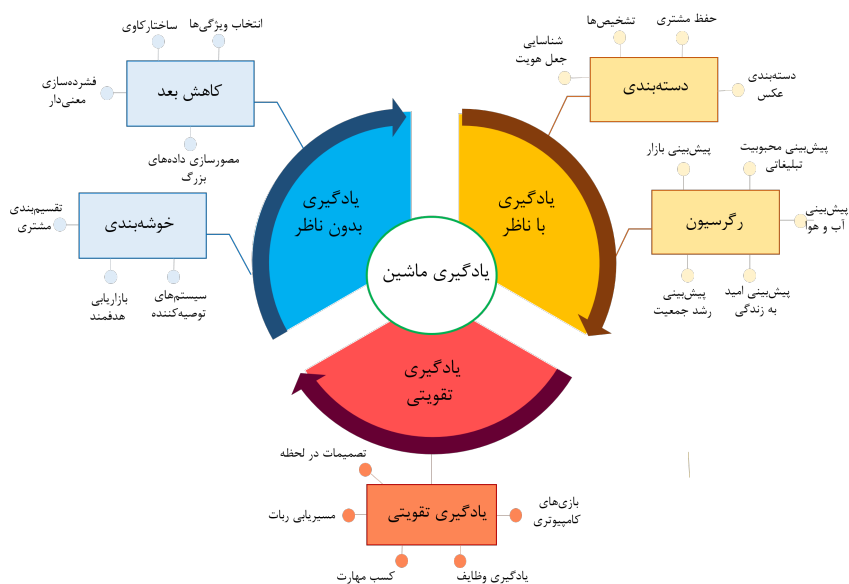
برای کسب اطلاعات بیشتر و تکمیلی در این زمینه، خواننده علاقمند را به [۴] ارجاع می‌دهیم.

۳ مروری بر درختان تصمیم

یکی از مسائلی که در یادگیری ماشین، تشخیص الگو و داده‌کاوی بررسی می‌شود، یادگیری با ناظر^۳ است. مسائل یادگیری با ناظر، با توجه به هدف، به دو دسته رگرسیون و دسته بندی یا طبقه‌بندی^۴ تقسیم می‌شوند (شکل ۳). مسائلی که متغیر هدف آن‌ها گسسته است، دسته‌بندی نامیده می‌شوند. دسته‌بندی از مسائل مهم در داده‌کاوی است که در آن، هدف، سازمان دهی و تخصیص داده‌ها به دسته‌ها یا طبقه‌های مجزا است. درختان تصمیم از محبوبترین روش‌های دسته‌بندی و پیش‌بینی محسوب می‌شوند و نمایانگر قوانینی هستند که به راحتی می‌توان آن‌ها را به زبان قابل فهم برای انسان درآورد (شکل ۴ نمونه‌ای از درخت تصمیم در تشخیص بیماری است). درخت تصمیم یکی از پرکاربردترین الگوریتم‌ها در بین الگوریتم‌های داده‌کاوی است. درخت تصمیم دقیقاً مانند یک درخت است، با این تفاوت که از ریشه به سمت پایین (برگ) رشد کرده است. از اجزای اصلی درخت تصمیم می‌توان، برگ، ریشه و شاخه را نام برد. برگ در حقیقت گره‌هایی است که

³Supervised Learning

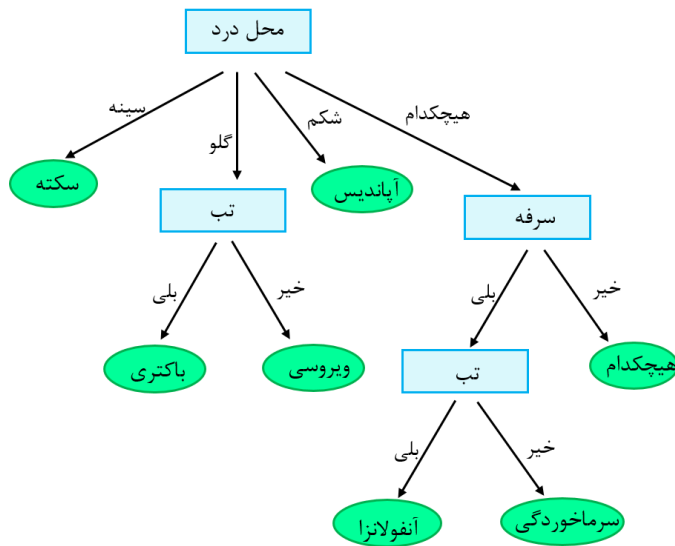
⁴Classification



شکل ۳: رویکردهای یادگیری ماشین

تقسیم‌های متوالی در آنجا پایان می‌یابد. برگ‌ها با یک دسته مشخص می‌شوند. منظور از ریشه، گره آغازین درخت است و در هر گره داخلی به تعداد جواب‌های ممکن شاخه ایجاد می‌شود. در الگوریتم درخت تصمیم، نمونه‌ها را دسته‌بندی می‌کنیم که در واقع دسته‌ها در انتهای گره‌های برگ قرار دارد. درخت تصمیم در مسائلی کاربرد دارد که بتوان آنها را به صورتی مطرح نمود که پاسخ واحدی به صورت نام یک دسته یا طبقه ارائه دهند. درخت تصمیم کاربردهای فراوانی در شناسایی الگوها، طبقه‌بندی و سیستم‌های پشتیبان تصمیم‌گیری و ... دارد. روش‌های ساخت درخت تصمیم معمولاً به صورت بالا به پایین عمل می‌کنند. یعنی در هنگام ساخت درخت، ابتدا ریشه درخت ساخته می‌شود، سپس هر یک از زیر شاخه‌ها به شاخه‌های دیگری تقسیم شده و این فرآیند تکرار می‌شود. نحوه انتخاب ویژگی برای تقسیم گره، روی اندازه و بازده درخت تولید شده، اثر مستقیم می‌گذارد. به طور کلی، تمامی الگوریتم‌های درخت تصمیم با تلاش در یافتن آن ویژگی شروع می‌شوند که بهترین تقسیم‌بندی نمونه‌ها را انجام دهد و اندازه درخت را حتی الامکان کوچک نمایند. از متداول‌ترین روش‌های انتخاب ویژگی، انتخاب ویژگی است که

⁵Top-Down Method



شکل ۴: نمونه‌ای از یک درخت تصمیم در تشخیص بیماری

بهره اطلاعاتی^۶ را بیشینه می‌کند. از دیگر معیارها، به شاخص جینی^۷، نسبت بهره^۸ و نظایر آن می‌توان اشاره کرد. الگوریتم‌های زیادی برای ساخت درخت تصمیم وجود دارند که از آن جمله می‌توان به الگوریتم‌های ID_3 ^۹ [۱۹]، $CART$ ^{۱۰}، $C4.5$ ^{۱۱} و... اشاره کرد. این الگوریتم‌ها با تقسیمات مکرر کار می‌کنند و در تقسیم فضای نمونه روش آنها مشابه است. از ویژگی‌های مهم آنها این است که همزمان با بهبود کیفیت تصمیم، اندازه درخت را نیز حداقل می‌کنند. در ادامه، به طور خلاصه به توضیح برخی از این الگوریتم‌ها می‌پردازیم.

۱.۳ الگوریتم ID_3

کوئینلن [۱۹] نخستین بار درختان تصمیم نمادین تحت نام ID_3 را معرفی کرد. این الگوریتم یکی از ساده‌ترین الگوریتم‌های درخت تصمیم است. در این الگوریتم درخت

⁶Information Gain

⁷Gini Index

⁸Gain ratio

⁹Iterative Dichotomiser

¹⁰Classification and Regression Tree

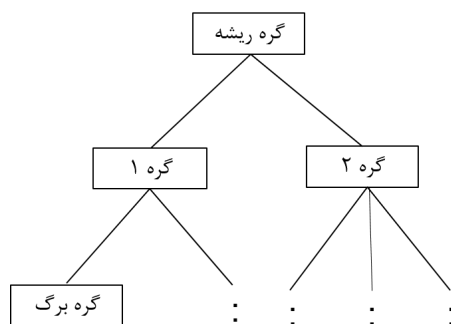
¹¹Classifier 4.5

تصمیم از بالا به پایین ساخته می‌شود. در الگوریتم ID^3 یک ویژگی یا صفت برای تقسیم نمونه‌ها انتخاب می‌شود. برای هر مقدار از ویژگی، یک شاخه ایجاد و زیرمجموعه متناسب نمونه‌هایی که مقدار ویژگی را دارند به وسیله شاخه تعیین می‌شود. ویژگی که دارای بیشترین بهره اطلاعاتی باشد، در گره ریشه درخت قرار می‌گیرد و این روند تا رسیدن به برگ ادامه می‌یابد (شکل ۵).

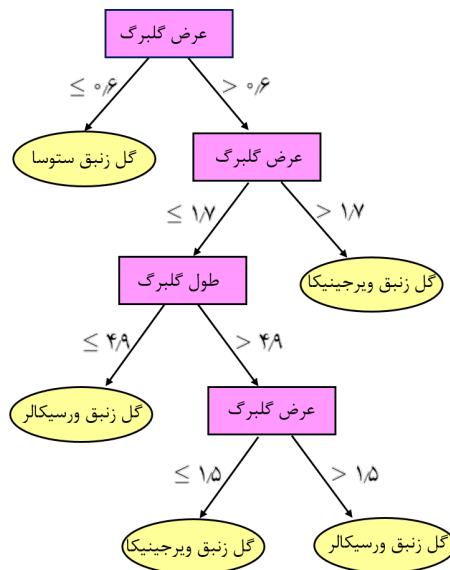
استنتاج از روی درخت تصمیم در ID^3 به این صورت انجام می‌شود که در هنگام تصمیم‌گیری برای یک نمونه جدید، ویژگی‌های آن نمونه با شروع از گره ریشه در برابر شرایط درخت مطابقت داده می‌شوند. چنانچه دقیقاً یک گره برگ وجود داشته باشد که شرایط آن (قرار گرفته بر روی مسیر از ریشه تا آن گره) برقرار شوند، به نمونه مورد نظر همان طبقه‌ای منتسب خواهد شد که به نمونه‌هایی که این شرایط را برآورده ساخته‌اند (آنهایی که در همین گره قرار دارند)، منتسب شده است. اگر حالتی غیر از این پیش بیاید، بایستی به نحو مناسبی در مورد آن تصمیم‌گیری شود.

۲.۳ الگوریتم $CART$

الگوریتم $CART$ به تقسیم‌بندی اولیه نیاز ندارد. شرایط گسترش درخت براساس آستانه‌ها (برای دامنه‌های پیوسته) که بطور پویا محاسبه می‌شوند، برقرار می‌شود. ویژگی مورد نظر در درخت می‌تواند بیش از یک بار و با آستانه‌های مختلف بکار رود ([۲]). الگوریتم $CART$ درختان دودویی می‌سازد و برای انتخاب گره‌های درخت، از شاخص جینی استفاده می‌کند. شکل ۶ نمونه‌ای از درخت $CART$ را نشان می‌دهد.



شکل ۵: ساختار درخت ID^3



شکل ۶: نمونه‌ای از یک درخت تصمیم CART

۳.۳ الگوریتم C4.5

الگوریتم C4.5 توسعه یافته الگوریتم ID3 است، که از معیار نسبت بهره برای تقسیم‌بندی گره‌ها در درخت استفاده می‌کند. از نقاط ضعف الگوریتم ID3 که در C4.5 رفع شده است به این موارد می‌توان اشاره کرد. الگوریتم C4.5 می‌تواند در مواردی که ویژگی‌ها دارای مقادیر گسسته یا پیوسته هستند، به کار رود. این الگوریتم قادر است با وجود مقادیر گمشده نیز درخت تصمیم خود را بسازد و این قابلیت را دارد که وزن‌های مختلف و غیر یکسانی را به برخی از ویژگی‌ها بدهد.

یک درخت تصمیم در صورتی که گره‌های زیادی داشته باشد دچار بیش برآزش^{۱۲} خواهد شد، به این معنی که خطای دسته‌بندی آن روی مجموعه داده‌ی آموزشی خیلی کمتر از خطای آن روی مجموعه داده‌ی آزمون خواهد بود. به عبارت دیگر، در حالت بیش برآزش، درخت تصمیم قابلیت تعمیم کمی خواهد داشت و دارای دقت کمتری است.

¹²Overfitting

۴.۳ هرس کردن درخت تصمیم

هرس،^{۱۳} یکی از روش‌های درخت تصمیم است که اندازه درختان تصمیم را با حذف قسمت‌هایی از درخت که قدرت کمی برای دسته‌بندی نمونه‌ها دارند، کاهش می‌دهد [۱۱]. در واقع هرس، پیچیدگی دسته‌بندی نهایی را کاهش می‌دهد، از این رو دقت پیش‌بینی را با کاهش بیش‌برازش بهبود می‌دهد. برای اجتناب از بیش‌برازش روش‌های پیش‌هرس^{۱۴} و پس‌هرس^{۱۵} وجود دارد. روش‌های پیش‌هرس با استفاده از معیارهای توقف، از رشد بیش از حد درخت جلوگیری می‌کنند. در حالی که روش‌های پس‌هرس، اجازه می‌دهند که درخت خیلی رشد کرده و دچار بیش‌برازش شود، سپس شاخه‌هایی که تاثیر زیادی در دقت دسته‌بندی ندارند را حذف کرده و اندازه درخت را کوچک می‌کند. هرس حداقل خطا^{۱۶} [۱۶]، هرس خطای بدبینانه^{۱۷} [۲۰]، هرس مبتنی بر خطا^{۱۸} [۲۰] هرس حداقل طول توصیف^{۱۹} [۱۴] و ... نمونه‌هایی از روش‌های هرس کردن هستند. روش‌های هرس کردن در مقالات مختلفی با هم مقایسه شده‌اند [۶، ۱۵]. بر اساس این مقایسه‌ها، روش هرس کردنی که در تمام شرایط بهتر از بقیه عمل کند، وجود ندارد.

در بسیاری از مسائل دسته‌بندی با داده‌های پیوسته سر و کار داریم، در حالی که برخی از روش‌های دسته‌بندی فقط با داده‌های گسسته کار می‌کنند. برای اینکه بتوان از این دسته‌بندها در مسائلی با داده‌های پیوسته استفاده کرد، لازم است که داده‌های پیوسته را به داده‌های گسسته تبدیل کرد. به فرآیند تبدیل داده‌های پیوسته به داده‌های گسسته، گسسته‌سازی^{۲۰} گفته می‌شود. در گسسته‌سازی دامنه تغییرات یک ویژگی به بازه‌های کوچک‌تر تقسیم می‌شود، سپس به هر یک از داده‌هایی که در یک بازه قرار دارند یک مقدار گسسته نسبت داده می‌شود. تعداد مقادیر ویژگی پس از گسسته‌سازی، برابر است با تعداد بازه‌هایی که در فرآیند گسسته‌سازی ایجاد شده است. بدیهی است یک گسسته‌سازی مناسب، تاثیر بسزایی در کیفیت دسته‌بندی دارد. از دیدگاه‌های مختلف می‌توان روش‌های

¹³Prun

¹⁴Pre Pruning

¹⁵Post Pruning

¹⁶Minimum Error Pruning

¹⁷Pessimistic Error Pruning

¹⁸Error Base Pruning

¹⁹Minimum Description Length

²⁰Discretization

گسسته‌سازی را تقسیم‌بندی کرد [۷]. روش‌های تقسیم‌کننده ^{۲۱} در مقابل روش‌های ادغام‌کننده ^{۲۲}، روش‌های سراسری ^{۲۳} در مقابل روش‌های محلی ^{۲۴}، روش‌های با ناظر در مقابل روش‌های بدون ناظر و روش‌های ایستا ^{۲۵} در مقابل روش‌های پویا ^{۲۶}.

برای مثال، روش‌های گسسته‌سازی تقسیم‌کننده به صورت بالا به پایین عمل می‌کنند. به این صورت که ابتدا برای تقسیم، بهترین نقطه برش را تعیین می‌کنند و به تدریج بازه و زیر بازه‌ها را تقسیم‌بندی می‌کنند. در مقابل روش‌های ادغام‌کننده به صورت پایین به بالا عمل می‌کنند به این صورت که ابتدا تمام مقادیر ویژگی مورد نظر را نقاط برش ممکن در نظر می‌گیرند و سپس در جستجوی حذف برخی از نقاط برش و ادغام بازه‌ها با یکدیگر، بازه‌های نهایی را تولید می‌کنند. روش‌های محلی برای تولید بازه‌ها به صورت محلی عمل می‌کنند، در حالیکه روش‌های سراسری برای تولید بازه‌ها از تمام داده‌ها استفاده می‌کنند. روش‌های ایستا از یک پارامتر مانند k برای تعیین حداکثر تعداد بازه‌ها برای هر ویژگی استفاده می‌کنند، در حالی که روش‌های پویا به دنبال k بازه بر روی تمام ویژگی‌ها هستند. گسسته‌سازی بدون ناظر، بدون در نظر گرفتن برچسب داده‌ها، آنها را گسسته‌سازی می‌کند در حالی که گسسته‌سازی با ناظر از برچسب داده‌ها نیز برای گسسته‌سازی استفاده می‌کند. برای مرور مقالاتی که در رابطه با گسسته‌سازی ارائه شده‌اند می‌توان به [۱۰] مراجعه کرد.

۴ درختان تصمیم فازی

اگر چه روش‌های زیادی برای ساخت درختان تصمیم در رابطه با داده‌های رده‌بندی یا پیوسته پیشنهاد شده‌اند، اما درختان تصمیم در برخی موارد قابل به کارگیری نیستند. یکی از معایب درخت تصمیم بی ثباتی آن می‌باشد و نسبت به داده‌های نویزدار بسیار حساس است. ساختار درخت تصمیم با اندکی نویز در مجموعه داده‌ها ممکن است به کلی تغییر کند و در مواردی که نیاز به تصمیم‌گیری عددی مورد نیاز است این نوع درخت تصمیم، غیر

²¹ Splitting

²² Merging

²³ Global

²⁴ Local

²⁵ Static

²⁶ Dynamic

قابل استفاده می‌شود. یک رهیافت مناسب در چنین مواردی، استفاده از درخت‌های تصمیم فازی (FDT^{۲۷}) است که در آن درخت تصمیم با استفاده از قوانین فازی ساخته می‌شود. درخت تصمیم فازی حساسیت کمتری نسبت به نویز دارد و علاوه بر این، از متغیرهای زبانی نیز پشتیبانی می‌کند. درخت تصمیم فازی یک سیستم استنتاج فازی است که قوانین آن ساختار درختی دارند. نحوه ساختن این درختان مشابه درخت تصمیم غیر فازی است که محاسبات آن با در نظر گرفتن یک درجه عضویت برای هر نمونه و به صورت فازی انجام می‌شود. در درخت تصمیم غیر فازی، مقادیر ورودی، یک مسیر مشخص از ریشه تا برگ را طی می‌کنند. اما در درخت تصمیم فازی، تمام برگ‌ها در دسته‌بندی نمونه مورد بررسی موثر هستند. میزان تاثیر هر برگ، به درجه عضویت نمونه مورد بررسی (با توجه به توابع عضویت مربوط به آن برگ) بستگی دارد. درختان تصمیم فازی اجازه می‌دهند که نمونه‌ها بتوانند چندین شاخه از درخت را با درجه‌های عضویت متفاوت دنبال کنند. علیرغم اینکه درخت تصمیم فازی در مقایسه با درخت تصمیم غیر فازی، دارای دقت بیشتر و قابلیت تفسیر بالاتری است، اما از نظر محاسباتی دارای سرعت کمتری است و نیاز به زمان بیشتری برای ساختن درخت دارد. درختان تصمیم فازی اجازه می‌دهند که نمونه‌ها بتوانند چندین تقسیم‌بندی را با درجه‌های عضویت متفاوت دنبال کنند. برای اجرای این مشخصات، درخت تصمیم فازی از ترم‌های زبانی فازی برای تعیین تقسیم‌بندی گره‌ها استفاده می‌کند. در درخت تصمیم فازی یک نمونه ممکن است در چندین برگ با درجه عضویت متفاوت قرار بگیرد. اگر چه از لحاظ محاسباتی، استنتاج درخت تصمیم فازی سرعت کمتری نسبت به استنتاج درخت تصمیم معمولی دارد، این هزینه برای داشتن دقت بالاتر همراه با قابلیت تفسیر بالا پرداخت می‌شود.

درخت تصمیم فازی شامل دو مرحله فرآیند ساخت و فرآیند استنتاج آن به منظور پیش بینی برچسب داده‌های جدید می‌باشد. دسته بزرگی از روش‌های استقرار درخت تصمیم فازی وجود دارند که بر پایه الگوریتم ID^۳ هستند که FID^۳ نامیده می‌شود. الگوریتم FID^۳ عبارت‌های زبانی فازی که از پیش تعریف شده‌اند را توسط مقادیر متغیر داده‌های آموزشی که فازی شده‌اند، به کار می‌گیرد. این روش از معیار بهره اطلاعاتی به منظور تعیین متغیر تقسیم‌بندی برای هر گره استفاده می‌کند. از جمله این موارد می‌توان به

²⁷Fuzzy Decision Tree

CART و *FUZZY - CART* و *UR - ID3* اشاره کرد. اولی روشی است که الگوریتم *CART* را برای ساخت درخت به کار می‌گیرد. در این روش الگوریتمی در قالب یک شبکه عصبی چند لایه برای یادگیری قوانین فازی به کار گرفته می‌شود که در نتیجه دانش تولید شده به وسیله *CART* اولیه را بهبود بخشیده، و قوانین فازی قابل درک تری تولید می‌کند. *UR - ID3* کار خود را با ساخت درخت تصمیم کلاسیک شروع کرده، و متعاقباً شرایط درخت را فازی‌سازی می‌کند. در کنار موارد فوق، دیگر روش‌های استقرای درخت تصمیم فازی که بر پایه روش استقرای *ID3* بنا نهاده شده‌اند را می‌توان در منابع [۲۶، ۸، ۱۲] مطالعه کرد. یک روش استقرای درخت تصمیم فازی را جانیکوف [۸] ارائه داد که بر اساس این روش درخت‌هایی تولید می‌شوند که می‌توانند هم برای دسته‌بندی و هم برای رگرسیون به کار گرفته شوند. در این روش، داده‌های آموزشی می‌توانند شامل مقادیر فازی باشند، اما عبارات‌های زبانی فازی به کار رفته در درخت می‌بایست به وسیله‌ی متخصصان از قبل تعریف شده باشد. لی و همکاران [۱۲] روشی پیشنهاد دادند که در آن ویژگی‌های با مقادیر گسسته بوسیله‌ی اعداد فازی تعیین می‌شوند. هر داده‌ی آموزشی، دارای اطلاعات هر طبقه به همراه درجه‌ی عضویت آن است. این روش استقرای درخت تصمیم فازی، توابع عضویت فازی برای ویژگی‌های پیوسته را به صورت خودکار می‌سازد. بعد از ساختن درخت، می‌توان عبارات‌های زبانی توصیفی را به توابع عضویت فازی تولید شده نسبت داد تا قابلیت درک درخت ارتقا پیدا کند.

وانگ و همکاران [۲۴]، مقاله‌ای تحت عنوان درختان تصمیم مبتنی بر قوانین فازی ارائه دادند. آنها یک شکل جدید از درخت تصمیم فازی را پیشنهاد دادند که بر خلاف درختان تصمیم رایج که تنها یک ویژگی در هر گره از درخت قرار داده می‌شود، هر گره از درخت، شامل یک قانون فازی است که چندین ویژگی را در بر دارد. استفاده از چندین ویژگی در یک گره، باعث شده است که اندازه درخت به حداقل برسد و گره‌هایی با خلوص بیشتری حاصل شوند. براساس یک مجموعه داده، درخت پیشنهادی آنها بهترین عملکرد را در مقایسه با درختان رایج، از نظر دقت و اندازه درخت تولید شده، داشته است.

برخی از متون روی ساخت توابع عضویت تمرکز کرده و تلاش کرده‌اند توابع عضویتی تولید کنند که منجر به ساخت درختان تصمیم بهتری شود. به عنوان مثال شیجائو و هونگ [۲۵]، یک ویژگی با مقادیر پیوسته را با استفاده از اعداد فازی و نظریه امکان،

گسسته‌سازی کرده‌اند. در مقاله آنها این واقعیت که انتخاب توابع عضویت در یک کلاس از توزیع‌های متقارن بر تولید درخت تصمیم تأثیر نمی‌گذارند، نیز ثابت شده است. پدریچ و سوسنوسکی [۱۸]، درخت تصمیم مبتنی بر خوشه‌بندی ارائه کردند که بر خلاف درخت‌های تصمیم متداول که در آن‌ها هر گره از درخت، یک ویژگی را شامل می‌شود، همه ویژگی‌ها در هر گره از درخت در نظر گرفته شده است. رشد درخت در این روش توسط گسترش گره‌ای از درخت که شامل داده‌هایی با بیشترین تنوع با توجه به برچسب آنها است، انجام شده است. در ساختار این درخت، ابتدا داده‌های آموزشی به C خوشه تقسیم شده‌اند و داده‌هایی که دارای شباهت بیشتری هستند در یک خوشه قرار گرفته‌اند. فرایند رشد درخت توسط یک معیار قطعی که میزان تنوع در هر خوشه است، ادامه می‌یابد. به عبارت دیگر، گره‌ای که بیشترین میزان تنوع را داراست برای خوشه‌بندی در مرحله بعدی درخت انتخاب شده و دارای C فرزند می‌گردد و خوشه‌های دیگر بدون فرزند باقی می‌مانند. فرایند رشد درخت تا زمانی که معیار تنوع صفر یا کمتر از یک مقدار آستانه مشخص شود، ادامه می‌یابد. زینل خوانی و افتخاری [۳۳]، الگوریتم‌های مختلفی را با استفاده از روش‌های گسسته‌سازی و تعریف توابع عضویت مختلف روی افرازاها معرفی کردند. علاوه بر این، یک الگوریتم جدید برای تولید توابع عضویت به نام افرازبندی فازی مبتنی بر آنتروپی فازی پیشنهاد کردند.

۱.۴ الگوریتم درخت تصمیم فازی

به‌طور کلی، درخت تصمیم فازی شامل دو جزء اصلی است: مرحله ساختن درخت تصمیم فازی و مرحله استنتاج درخت برای پیش‌بینی نمونه‌های جدید. ابتدا برخی از نمادها و روابط استفاده شده در الگوریتم فازی را معرفی می‌کنیم.

- مجموعه ویژگی‌های داده به صورت $\{A_1, A_2, \dots, A_k, Y\}$ تعریف می‌شوند که Y متغیر هدف، k تعداد ویژگی داده‌ها و A_i ، i امین ویژگی است.
- مجموعه‌ای از مقادیر متغیر هدف $\{c_1, c_2, \dots, c_m\}$ است که $Y \in \{c_1, c_2, \dots, c_m\}$ تعداد m کلاس‌ها یا دسته‌های موجود در یک مجموعه داده می‌باشد.
- مجموعه داده فازی S توسط $S = \{(X_1, \mu_S(X_1)), (X_2, \mu_S(X_2)), \dots, (X_n, \mu_S(X_n))\}$

نشان داده می‌شوند که X_i ، i امین نمونه و $\mu_S(X_i)$ درجه عضویت X_i در S می‌باشد و n تعداد نمونه در یک مجموعه داده می‌باشد. X_i برداری است که شامل ویژگی‌ها و متغیر هدف می‌باشد.

• i امین نمونه توسط $\mathbf{X}_i = [x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(k)}, y_i]^T$ نشان داده می‌شود که $x_i^{(j)}$ مقدار j امین ویژگی این نمونه است و y_i مقدار متغیر هدف آن می‌باشد.

• ترم‌های فازی تعریف شده روی i امین ویژگی، A_i ، به صورت $\{F_i^{(1)}, F_i^{(2)}, \dots, F_i^{(k)}\}$ نشان داده می‌شوند که $F_i^{(j)}$ ، j امین ترم فازی این ویژگی است و r_i تعداد ترم‌های فازی تعریف شده روی ویژگی A_i می‌باشد.

• تابع عضویت مربوط به ترم فازی $F_i^{(j)}$ به صورت $\mu_{F_i^{(j)}}$ نشان داده می‌شود.

• تعداد نمونه‌های مجموعه داده فازی S توسط $|S|$ نشان داده می‌شود و به صورت زیر تعریف می‌شود:

$$|S| = \sum_{i=1}^n \mu_S(X_i)$$

• نمونه‌های مجموعه داده‌ی فازی S که به کلاس i تعلق دارند توسط $S_{y=c_i}$ نمایش داده می‌شوند.

• اگر S مجموعه داده فازی یک گره را نشان دهد، $S[F_i^{(j)}]$ مجموعه داده‌ی فازی گره فرزند مربوط به ترم فازی $F_i^{(j)}$ را نشان می‌دهد.

فرض کنید S مجموعه داده فازی گره والد باشد و A_i متغیر تقسیم‌بندی با نرم‌های فازی $\{F_i^{(1)}, F_i^{(2)}, \dots, F_i^{(r_i)}\}$ باشد و مجموعه داده فازی گره برگ مربوط به ترم فازی $F_i^{(j)}$ ، $S[F_i^{(j)}]$ باشد. درجه عضویت h امین نمونه $\mathbf{X}_h = [x_h^{(1)}, x_h^{(2)}, \dots, x_h^{(k)}, y_h]^T$ در $S[F_i^{(j)}]$ به صورت زیر می‌باشد [۳۳]

$$\mu_{S[F_i^{(j)}]}(\mathbf{X}_h) = \mu_S(\mathbf{X}_h) \times \mu_{F_i^{(j)}}(x_h^{(i)}) \quad (۲)$$

که $\mu_S(\mathbf{X}_h)$ در عضویت $\mathbf{X}_h$ در S و $\mu_{F_i^{(j)}}(x_h^{(i)})$ درجه عضویت $x_h^{(i)}$ در تابع عضویت مربوط به ترم فازی $F_i^{(j)}$ می باشد که $\mu_{F_i^{(j)}}$ نامیده می شود. به منظور تعمیم دادن آن هر عملگر $t - norm$ می تواند جایگزین عملگر ضرب شود.

الگوریتم $FID3$ ، متغیری که حداکثر مقدار بهره اطلاعاتی فازی $^{28} (FIG)$ را دارا باشد، به عنوان متغیر تقسیم بندی انتخاب می کند. در محاسبه FIG از آنتروپی فازی $^{29} (FE)$ استفاده می شود که به صورت زیر تعریف می شود.

$$FE(S) = \sum_{i=1}^m - \frac{|S_{y=c_i}|}{|S|} \log_2 \frac{|S_{y=c_i}|}{|S|} \quad (3)$$

معیار FIG متغیر A_i مربوط به مجموعه داده ی فازی S به صورت زیر تعریف می شود:

$$FIG(S, A_i) = FE(S) - \sum_{j=1}^{r_i} w_j \times FE(S[F_i^{(j)}]) \quad (4)$$

که $EF(S)$ آنتروپی فازی مجموعه داده فازی S می باشد و $FE(S[F_i^{(j)}])$ آنتروپی فازی j امین گره فرزند است و w_j کسری از نمونه هایی است که به j امین گره فرزند تعلق دارند. به صورت زیر تعریف می شود

$$w_j = \frac{|S[F_i^{(j)}]|}{\sum_{k=1}^{r_i} |S[F_i^{(k)}]|} \quad (5)$$

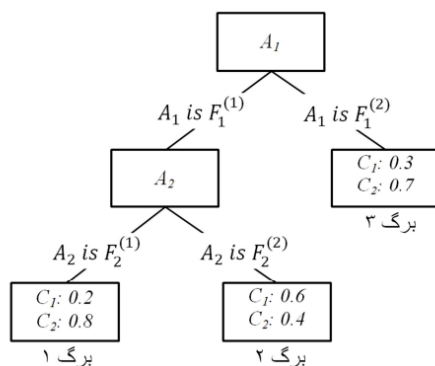
اولین ترم در رابطه (4) فقط آنتروپی فازی S است و دومین ترم مقدار کاهش یافته ی آنتروپی فازی پس از تقسیم شدن S توسط متغیر A_i می باشد. بنابراین $FIG(S, A_i)$ مقدار کاهش یافته آنتروپی فازی به دست آمده توسط مقادیر متغیر A_i می باشد. برخی دیگر از مطالعات انجام شده، روی بهبود روش ساخت درخت تصمیم تمرکز کرده اند. بعضی از این مطالعات برای بهبود روش ساخت درخت، یک معیار جداسازی جدیدی پیشنهاد کرده اند. یوان و شاو [۲۶] الگوریتمی را پیشنهاد دادند که در هر گره، ویژگی ای انتخاب می شود که حداقل ابهام را در دسته بندی ایجاد کند. جنسن و شن [۹] از معیار

²⁸Fuzzy Information Gain

²⁹Fuzzy Entropy

جداسازی مبتنی بر نظریه مجموعه‌های سخت فازی برای ساخت درخت تصمیم فازی استفاده کردند. وانگ و بورگل [۲۲] معیارهای جداسازی بهره اطلاعاتی و نسبت بهره را در درختان تصمیم فازی توسعه دادند و روشی را برای برخورد با مقادیر گمشته ارائه کردند. بعضی از مطالعات به بهینه‌سازی درخت تصمیم فازی پرداخته‌اند. وانگ و همکاران [۲۳] به بررسی نقص‌های FID_3 پرداختند و روش بهینه‌سازی برای درخت تصمیم بر مبنای حداقل کردن تعداد برگ‌ها و اندازه درخت پیشنهاد دادند. بعلاوه نشان دادند که ساخت کوچک‌ترین درخت تصمیم فازی یک مسئله NP - سخت است.

اولارو و وهنکل [۱۷] روش جدیدی از درختان تصمیم فازی، به نام درختان تصمیم نرم^{۳۰} ارائه دادند که ترکیبی از رشد و هرس درخت است. علاوه بر این روش‌هایی را برای بهبود تعمیم درخت تصمیم نرم پیشنهاد دادند.



شکل ۷: نمای کلی از یک درخت تصمیم فازی

۵ مثال عددی

همانطور که ذکر شد درخت تصمیم فازی، توسعه یافته درخت تصمیم کلاسیک برای استخراج اطلاعات از داده‌های نویزدار و با طبقه‌بندی نامشخص است. این تئوری، مجموعه‌های فازی را برای نمایش مجموعه داده‌ها اعمال می‌کند و همزمان رشد و هرس درخت را برای تعیین ساختار درخت به کار می‌گیرد. در این قسمت برای روشن‌تر شدن

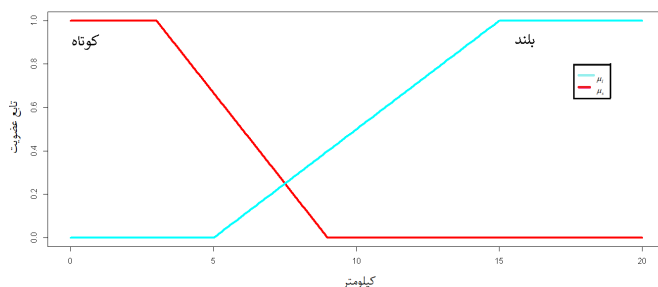
³⁰ Soft Decision Trees

بحث به ذکر یک مثال عددی می‌پردازیم. فرض کنید تراکم ترافیک به صورت یک عدد فازی با تابع عضویت زیر معرفی شود ([۳۰]):

$$\mu_{sh}(X) = \begin{cases} 1 & x < 3 \\ 1.5 - \frac{x}{6} & 3 \leq x < 9 \\ 0 & x \geq 9 \end{cases} \quad (6)$$

$$\mu_l(X) = \begin{cases} 0 & x < 5 \\ \frac{x}{10} - 0.5 & 5 \leq x < 15 \\ 1 & x \geq 15 \end{cases} \quad (7)$$

همانطور که در شکل ۸ نیز مشخص است، هر ترافیک می‌تواند در مجموعه بلندی و



شکل ۸: تابع عضویت تراکم ترافیک

کوتاهی قرار داده شود. برای مثال اگر ترافیک ۳ کیلومتر باشد، تابع عضویت بلندی آن صفر و تابع عضویت کوتاهی برای آن یک است. همچنین یک ترافیک با طول ۶ کیلومتر با اندازه ۰/۵ به مجموعه کوتاه و با اندازه ۰/۱ به مجموعه بلند تعلق دارد. در مرحله بعد، آنتروپی فازی و بهره اطلاعاتی مجموعه داده‌های فازی را برای ساختن درخت محاسبه می‌کنیم.

آنتروپی در الگوریتم $ID3$ با استفاده از رابطه (۳) محاسبه می‌شود و برابر است با

$$FE(S) = 0.940$$

و بهره اطلاعاتی (۴) به صورت زیر به دست می‌آید:

$$FE(S[l]) = - \left(\frac{0.96}{2.09} \right) \log_2 \left(\frac{0.96}{2.09} \right) - \left(\frac{1.13}{2.09} \right) \log_2 \left(\frac{1.13}{2.09} \right) = 0.995$$

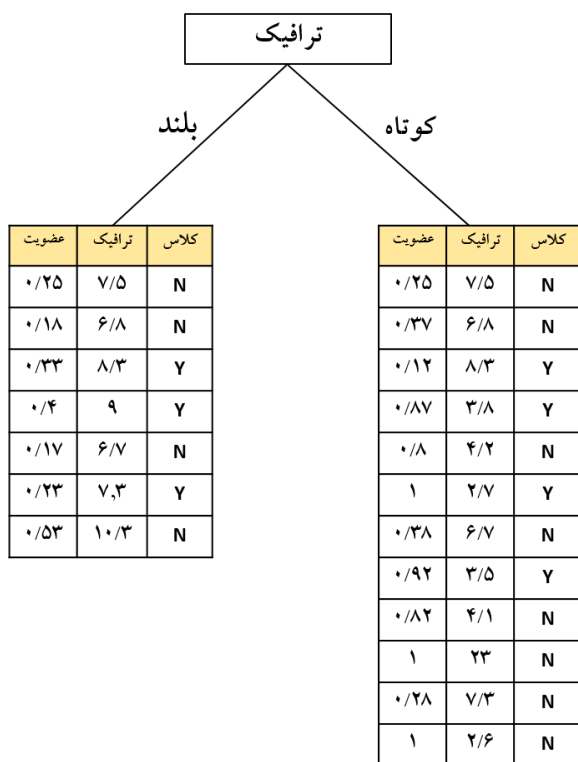
$$FE(S[sh]) = - \left(\frac{6.01}{7.81} \right) \log_2 \left(\frac{6.01}{7.81} \right) - \left(\frac{1.80}{7.81} \right) \log_2 \left(\frac{1.80}{7.81} \right) = 0.779$$

$$FIG(S) = 0.940 - \left(\frac{2.09}{9.90} \right) 0.995 - \left(\frac{7.81}{9.90} \right) 0.779 = 0.1154$$

با استفاده از این داده‌ها یک درخت تصمیم فازی طراحی می‌کنیم. ابتدا یک گره ریشه ایجاد کرده که دارای مجموعه‌ای فازی از تمام داده‌ها با مقدار عضویت ۱ باشد، سپس از توابع عضویت تراکم ترافیک در روابط (۶) و (۷) برای ساختن درخت استفاده می‌کنیم. با استفاده از داده‌ها دو گره فرعی ایجاد می‌کنیم، که در آن مقادیر عضویت در این گره‌های فرعی حاصل ضرب مقادیر عضویت اصلی در ریشه و مقادیر عضویت تراکم ترافیک است. به عنوان مثال، برای گره سمت چپ که مربوط به ویژگی بلند بودن ترافیک است، مقدار عضویت داده اول برابر با ۰/۲۵ است. مقدار عضویت جدید این داده در این گره ۰/۲۵ است. که با استفاده از فرمول (۲) به صورت زیر محاسبه می‌شود:

$$\mu_{new} = 0.25 * 1 = 0.25$$

حال باید نسبت کلاس C_k را محاسبه کنیم. این نسبت برابر است با مجموع مقادیر عضویت کلاس C_k به مجموع همه مقادیر عضویت. به عنوان مثال، در گره فرعی سمت چپ، نسبت کلاس N با ۷ داده، برابر ۵۴ درصد است. پس از آن، نسبت و تعداد مجموعه داده را با حدود کنترل فازی (که در این مثال ۰/۸۵ و ۰/۷۵ در نظر گرفته شده است) مقایسه می‌کنیم. اگر آنها کوچکتر از حدود کنترل باشند و اگر ویژگی‌هایی برای طبقه‌بندی نیز وجود داشته باشد، ادامه می‌دهیم. برای ایجاد یک گره جدید، این فرآیندها را تا زمانی که شرایط توقف برآورده شود، تکرار می‌کنیم. به عنوان مثال، نسبت کلاس Y



شکل ۹: تقسیم درخت در الگوریتم FID_3

در گره فرعی سمت راست ۷۷ درصد است. اگر پارامتر کنترل فازی تعریف شده توسط کاربر ۷۰ درصد باشد، گسترش این گره را متوقف می‌کنیم. که این بدین معنی است که اگر ترافیک کوتاه باشد، احتمال عدم رانندگی و رانندگی خوب به ترتیب ۲۳ درصد و ۷۷ درصد است. عضویت هر کلاس به صورت زیر محاسبه می‌شود: گره چپ:

$$C_N = 0.25 + 0.18 + 0.17 + 0.53 = 1.13$$

$$C_Y = 0.33 + 0.4 + 0.23 = 0.96$$

$$\text{مقدار کل عضویت} : 1.13 + 0.96 = 2.09$$

$$N \text{ نسبت کلاس} : \frac{1.13}{2.09} = 0.54 \quad Y \text{ نسبت کلاس} : \frac{0.96}{2.09} = 0.46$$

گره راست:

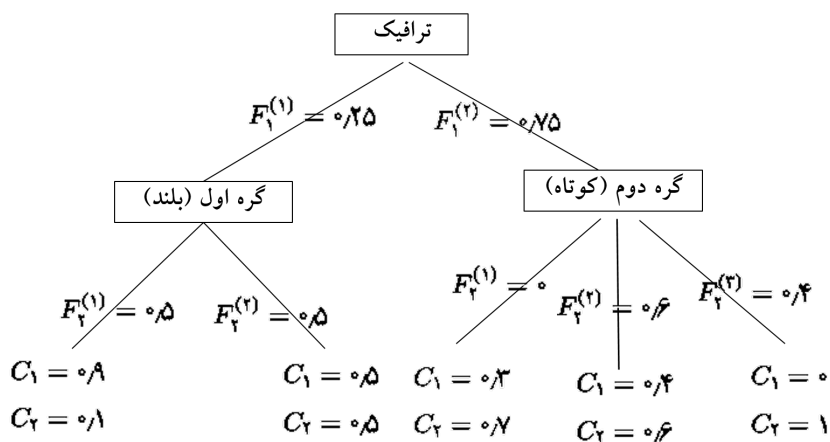
$$C_N = 0.25 + 0.37 + 0.8 + 0.38 = 1.8$$

$$C_Y = 0.12 + 0.87 + 1.0 + 0.92 + 0.82 + 1.0 + 0.28 + 1.0 = 6.01$$

$$\text{مقدار کل عضویت} : 1.8 + 6.01 = 7.81$$

$$N \text{ نسبت کلاس} : \frac{1.8}{7.81} = 0.23 \quad Y \text{ نسبت کلاس} : \frac{6.01}{7.81} = 0.77$$

شکل ۱۰ قواعد استنتاج فازی این داده‌ها را ارائه می‌دهد. گره درخت و گره برگ به



شکل ۱۰: نمایش درخت قواعد استنتاج فازی

ترتیب هر کدام نشان دهنده مقدار تابع عضویت ویژگی در گره و نسبت هر کلاس در گره است. با استفاده از روش $\times - \times - +$ امانول و همکاران [۲۱] نتیجه می‌شود که نمونه به ترتیب با احتمال ۰.۳۵۵ و ۰.۶۴۵ متعلق به کلاس C_1 و C_2 است که این احتمالات مکمل هستند. در زیر محاسبات C_1 و C_2 ارائه شده است.

$$C_1 = 0.25 \times 0.5 \times 0.8 + 0.25 \times 0.5 \times 0.5 + 0.75 \times 0 \times 0.3 + 0.75 \times 0.6 \times 0.4 + 0.75 \times 0.4 \times 0 = 0.355$$

$$C_2 = 0.25 \times 0.5 \times 0.1 + 0.25 \times 0.5 \times 0.5 + 0.75 \times 0 \times 0.7 + 0.75 \times 0.6 \times 0.6 + 0.75 \times 0.4 \times 1 = 0.645$$

$$C_1 + C_2 = 1$$

۶ کاربردهای مختلف از درختان تصمیم فازی

در این بخش از مقاله، با توجه به اهمیت مطالب ارائه شده و نقش درخت تصمیم فازی در سایر علوم و صنایع مختلف به بیان کاربردهای متفاوتی از این الگوریتم‌ها در علوم مختلف می‌پردازیم. واضح است که مثال‌ها و نمونه‌های مطرح شده در این قسمت، بخش کوچکی از کاربرد درخت تصمیم در علوم مختلف است که به ذکر آن پرداخته‌ایم و بیان کلیه کاربردها خارج از وسع این مقاله است.

رضایی و همکاران [۳] در مقاله‌ای با عنوان با استفاده از داده‌های مربوط به تعیین مزاج افراد و با استفاده از درخت تصمیم فازی با الگوریتم ID_3 برای مدل‌سازی و تشخیص مزاج افراد استفاده شده است. آنها دو درخت تصمیم‌گیری یکی برای گرمی، سردی و دیگری برای تری، خشکی ساختند و با استفاده از این دو درخت تقسیمات مزاجی را انجام دادند. نتایج حاصل مورد تایید پزشکان متخصص طب سنتی ایرانی قرار گرفت.

تجزیه و تحلیل سیگنال EEG یک مشکل مرتبط در انفورماتیک سلامت است و توسعه آن می‌تواند در تشخیص تشنج یا بیماری صرع کمک کند. رابکان و همکاران [۲۷] درخت تصمیم فازی را به عنوان طبقه‌بندی فازی برای تشخیص تشنج بیماران صرع استفاده کردند و نشان دادند در مقایسه با سایر مطالعات، این روش برای تشخیص تشنج یا بیماری صرع، بسیار موثر است.

ادریس و اسماعیل [۲۹] یک درخت تصمیم فازی را به عنوان روش طبقه‌بندی در تشخیص سرطان پستان پیشنهاد کردند و نتایج را با استفاده از سه مجموعه داده سرطان پستان تجزیه و تحلیل کرده و نشان دادند که ترکیب الگوریتم FID_3 با روش FUZZYDBD از کارایی بالایی برخوردار است.

مطالعات حاصل از پیش‌بینی امواج بعنوان اساس طراحی‌های مهندسی دریا، نقش مهمی در ساخت، نگهداری و مدیریت پروژه‌های عمرانی ساحلی و فراساحلی، دارد. امواج ناشی از باد، به دلیل دارا بودن بیشترین انرژی و فراوانی، از مهمترین امواج در دریا محسوب میشوند. بزرگ‌زاده و همکاران [۱] در تحقیق خود، از سیستم استنتاج تطبیقی عصبی- فازی، برای پیش‌بینی خصوصیات امواج استفاده کردند و روش‌های مختلف را با هم مقایسه نمودند.

رابکان و همکاران [۲۸] با استفاده از درخت تصمیم فازی، الگوریتمی را برای تشخیص

پره‌های معیوب و سالم موتور هواپیما پیشنهاد دادند که در مقایسه با سایر الگوریتم‌ها، از دقت بالایی برخوردار است.

لودویگ و همکاران [۲۱] درخت تصمیم فازی را برای طبقه‌بندی داده‌های بیان ژن به کار بردند و پنج مجموعه داده سرطان با ابعاد بالا را مورد تجزیه و تحلیل قرار دادند. آنها نشان دادند که درخت تصمیم فازی از نظر دقت، در چهار مورد از پنج مجموعه داده، دارای دقت بالایی است.

امانول و همکاران [۲۱] نسخه جدیدی از الگوریتم ID_3 را برای تولید درخت تصمیم فازی پیشنهاد کردند و از الگوریتم پیشنهادی خود برای تشخیص پتانسیل ترانسفورماتورها با تجزیه گاز در نفت استفاده کردند.

لوانشکو و زیاتسوا [۱۳] در مقاله‌ای با عنوان ”درختان تصمیم‌گیری فازی در سیستم پشتیبانی تصمیم‌گیری پزشکی” از درخت تصمیم فازی برای تشخیص بیماری سرطان پستان استفاده کردند.

۷ نتیجه‌گیری

اگر داده‌ها دارای نواقص و مشکلاتی چون اغتشاش، حجم کم نمونه، دقت کم و ... باشند، درخت تصمیم کارایی نداشته و برای رفع این مشکل از درخت تصمیم فازی استفاده می‌شود. در این روش درخت تصمیم با استفاده از قوانین فازی ساخته می‌شود. درخت تصمیم فازی یک سیستم استنتاجی فازی است که قوانین آن ساختار درختی دارند. این درختان به نمونه‌ها اجازه می‌دهند که بتوانند چندین تقسیم بندی را با درجه عضویت‌های متفاوت دنبال کنند. این درختان مشابه درخت تصمیم معمولی شامل دو فرآیند ساخت و استنتاج می‌باشند. فرایند ساخت درخت تصمیم فازی در اکثر مواقع به تقلید از درخت تصمیم معمولی است که چند روش کلی در متن مفصل معرفی شده است، ضمن اینکه مثال و کاربرد های مختلفی از درختان تصمیم فازی در مقاله ذکر گردیده است. با توجه به مباحث مطرح شده در مقاله این روش را می‌توان در مواجهه با بسیاری از داده‌ها مورد استفاده قرار داد.

مراجع

- [۱] بزرگ زاده لیاقت، بختیاری مرتضی، شهنی کرم زاده نیما، اسماعیل دوست محمد. تخمین ارتفاع امواج ناشی از باد با استفاده از سیستم استنتاج تطبیقی عصبی-فازی، درخت تصمیم و روشهای تجربی در بندر بوشهر. (۱۳۹۷). علوم مهندسی و آبیاری (مجله علوم کشاورزی). دوره ۴۱، شماره ۲، ۱۶۶-۱۵۱.
- [۲] جعفری، بهزاد، عادل آذر. درخت تصمیم فازی؛ رویکردی نوین در تدوین استراتژی. (۱۳۹۲). پژوهش‌های مدیریت عمومی. سال ششم شماره ۱۹، ۳۹-۲۵.
- [۳] رضایی اینانلو، سلیمانان قره چیق، و مجاهدی. (۱۳۹۷). ارائه مدلی بر مبنای درخت تصمیم فازی جهت تشخیص مزاج در طب سنتی ایرانی. مجله طب سنتی اسلام و ایران، ۹(۳)، ۲۲۷-۲۳۷.
- [۴] طاهری، محمود، ماشین چی، ماشالله. مقدمه‌ای بر احتمال و آمار فازی. (۱۳۹۲). انتشارات دانشگاه شهید باهنر کرمان.
- [5] Cintra, M. E., Monard, M. C., & Camargo, H. A. (2013). A fuzzy decision tree algorithm based on c4. 5. Hot Topics.
- [6] Esposito, F., Malerba, D., Semeraro, G., & Kay, J. (1997). A comparative analysis of methods for pruning decision trees. IEEE transactions on pattern analysis and machine intelligence, 19(5), 476-491.
- [7] Gupta, A., Mehrotra, K. G., & Mohan, C. (2010). A clustering-based discretization for supervised learning. Statistics & probability letters, 80(9-10), 816-824.
- [8] Janikow, C. Z. (1998). Fuzzy decision trees: issues and methods. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 28(1), 1-14.
- [9] Jensen, R., & Shen, Q. (2005). Fuzzy-rough feature significance for fuzzy decision trees. In Proc. 2005 UK Workshop on Computational Intelligence, 89-96.

- [10] Kotsiantis, S., & Kanellopoulos, D. (2006). Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, 32(1), 47-58.
- [11] Kuncheva, L. I. (2014). *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.
- [12] Lee, K. M., Lee, K. M., Lee, J. H., & Lee-Kwang, H. (1999). A fuzzy decision tree induction method for fuzzy data, in *Proc. of 8th IEEE Inter. conference on Fuzzy Systems*, 16 -21.
- [13] Levashenko, V., & Zaitseva, E. (2012). Fuzzy decision trees in medical decision making support system. In *2012 Federated Conference on Computer Science and Information Systems (FedCSIS)* (pp. 213-219). IEEE.
- [14] Mehta, M., Rissanen, J., & Agrawal, R. (1995). MDL-Based Decision Tree Pruning. In *KDD*, 21(2), 216-221.
- [15] Mingers, J. (1989). An empirical comparison of pruning methods for decision tree induction. *Machine learning*, 4(2), 227-243.
- [16] Niblett, T., & Bratko, I. (1987). Learning decision rules in noisy domains. In *Proceedings of Expert Systems' 86, the 6th Annual Technical Conference on Research and development in expert systems III*, 25-34.
- [17] Olaru, C., & Wehenkel, L. (2003). A complete fuzzy decision tree technique. *Fuzzy sets and systems*, 138(2), 221-254.
- [18] Pedrycz, W., & Sosnowski, Z. A. (2005). C-fuzzy decision trees. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4), 498-511.
- [19] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1), 81-106.

- [20] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann.
- [21] Umamori, M., Okamoto, H., Hatono, I., Tamura, H., Kawachi, F., Umedzu, S., Kinoshita, J. Fuzzy Decision Trees by Fuzzy ID3 algorithm and Its Application to Diagnosis Systems. In Proceedings of the Third IEEE Conference on Fuzzy Systems, vol. 3, pp. 2113-2118, Orlando, June, 1994.
- [22] Wang, X., & Borgelt, C. (2004). Information measures in fuzzy decision trees. In 2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No. 04CH37542), 85-90.
- [23] Wang, X., Chen, B., Qian, G., & Ye, F. (2000). On the optimization of fuzzy decision trees. Fuzzy Sets and Systems, 112(1), 117-125.
- [24] Wang, X., Liu, X., Pedrycz, W., & Zhang, L. (2015). Fuzzy rule based decision trees. Pattern Recognition, 48(1), 50-59.
- [25] Xizhao, W., & Hong, J. (1998). On the handling of fuzziness for continuous-valued attributes in decision tree generation. Fuzzy Sets and Systems, 99(3), 283-290.
- [26] Yuan, Y., & Shaw, M. J. (1995). Induction of fuzzy decision trees. Fuzzy Sets and systems, 69(2), 125-139.
- [27] Rabcan, J., Levashenko, V., Zaitseva, E., & Kvassay, M. (2020). Review of methods for EEG signal classification and development of new fuzzy classification-based approach. IEEE Access, 8, 189720-189734.
- [28] Rabcan, J., Levashenko, V., Zaitseva, E., Kvassay, M., & Subbotin, S. (2019). Application of fuzzy decision tree for signal classification. IEEE Transactions on Industrial Informatics, 15(10), 5425-5434.

- [29] Idris, N. F., & Ismail, M. A. (2021). Breast cancer disease classification using fuzzy-ID3 algorithm with FUZZYDBD method: automatic fuzzy database definition. *PeerJ Computer Science*, 7, e427.
- [30] Liang, G., van den Berg, J., & Kaymak, U. (2005). A comparative study of three Decision Tree algorithms: ID3, Fuzzy ID3 and Probabilistic Fuzzy ID3. Netherlands: Erasmus University.
- [31] Ludwig, S. A., Picek, S., & Jakobovic, D. (2018). Classification of cancer data: analyzing gene expression data using a fuzzy decision tree algorithm. In *Operations research applications in health care management* (pp. 327-347). Springer, Cham.
- [32] Zadeh, L. A. (1965). Fuzzy Sets, *Inform. Control*, 8: 338-353.
- [33] Zeinalkhani, M., & Eftekhari, M. (2014). Fuzzy partitioning of continuous attributes through discretization methods to construct fuzzy decision tree classifiers. *Information Sciences*, 278, 715-735.