پنجمین کنفرانس سیستمهای فازی ایران
هفدهم الی نوزدهم شهریورماه ۱۳۸۳

بهبود یادگیری درخت تصمیم با استفاده از منطق فازی و هرس کردن آن

ستار هاشمی[۱] ، کارو لوکس[۲] ، محمدرضا کنگاوری[۳]
دانشکده مهندسی کامپیوتر
دانشگاه علم و صنعت ایران
S_hashemi@iust.ac.ir

چکیده

درخت تصمیم از جمله مهمترین روشهای یادگیری از نمونه ها است که طی سالهای گذشته مورد توجه بوده است. از جمله نقاط ضعف درخت تصمیم، رشد بی رویه آن در محیط های نویزی است که منجر به کاهش جامعیت و افت دقت یادگیری آن می شود. برای مقابله با کاهش دقت یادگیری در محیط نویزی می توان از دو متد منطق فازی و هرس کردن در کنار درخت تصمیم استفاده کرد. از ویژگیهای درخت تصمیم فازی در مقایسه با درخت تصمیم هرس شده ، قابلیت تقسیم نرم و از جمله معایب آن، نیاز به داشتن متغیرهای فازی و مجموعه های آموزشی دارای درجه عضویت می‌باشد. نتایج نشان میدهد که در صورت داشتن مجموعه های آموزشی دارای توابع عضویت، درخت تصمیم فازی نسبت به روشهای هرس درخت تصمیم دقت بیشتری خواهد داشت.

مقدمه

درخت های تصمیم در واقع یک تخمین زننده احتمالی هستند که با استفاده از هیوریستیک های مختلف مبتنی بر تئوری اطلاعات و یا مبتنی بر نظریه آمار توسعه داده می شوند[1,3]. یکی از مشکلات درخت تصمیم ، قراردادن داده ها روی هم همدیگر (overfitting) ناشی ازداده های ناکامل می باشد. عدم جامعیت داده ها می تواند ناشی از نویز ، خطای اندازه‌گیری، ارزیابی موضوعی ، ضعف زبان توصیف یا بطور ساده ناشی از داده های گزارش نشده باشد. برای برخورد با این مشکلات میتوان از تکنیک های هرس درخت تصمیم و نیز درختهای تصمیم فازی بهره برد .

هرس کردن یکی از راههای مبارزه با عدم قطعیت موجود در داده ها می باشد که بطور کلی به دو دسته تقسیم میشود ۱- روش هرس هنگام توسعه درخت یا پیش هرس ۲- هرس بعد از اتمام توسعه درخت یا پس‌هرس که هر کدام ویژگیهای خاص خود را دارند. روش پیش هرس بصورت عمومی یک مقدار آستانه را برای معیار انتخاب درنظر می گیرد که اگر مقدار محاسبه شده از آن سطح بیشتر باشد تقسیم را ادامه داده و در غیر اینصورت درخت را هرس می کند. روش پس هرس برای ایجاد تعادل بین دقت و پیچیدگی درخت پیشنهاد شده است که درخت را پس از توسعه کامل هرس می کند و از محاسن آن میتوان به استفاده از کلیه اجزاء اطلاعات برای ساخت درخت با دقت مطلوب اشاره کرد اگر چه هزینه بیشتری نسبت به روش قبل دارد.

۱- دانشجوی دکتری کامپیوتر دانشگاه علم و صنعت ایران
۲- استاد گروه برق و کامپیوتر دانشگاه تهران
۳- استادیار دانشکده کامپیوتر دانشگاه علم و صنعت ایران

رویکرد دوم یعنی رویکرد فازی در واقع ترکیبی از درخت تصمیم و منطق فازی می باشد. منطق و مجموعه فازی اجازه مدل کردن عدم قطعیت مربوط به زبان را به ما میدهد یعنی چهارچوبی نمادی برای دانش ارائه میدهد که به جامعیت دانش کمک کرده و میتواند دانش را با جزئیات دقیق تر مدل کند. نمایش فازی بطور عمومی در مسائلی مانند عدم قطعیت ، نویز ، داده های غیر صحیح بکار گرفته شده است و نتایج موفقیت آمیزی از خود نشان داده است. در این مقاله دو متد برای توسعه و استنتاج درخت تصمیم فازی در مقایسه با دو روش متداول هرس درخت تصمیم مورد بررسی قرار می‌گیرد.

هرس درخت تصمیم

کم‌رنگ شدن هزینه محاسباتی بدلیل وجود کامپیوترهای قدرتمندتر از یک طرف و کارایی بیشتر روشهای مختلف پس‌هرس از طرف دیگر موجب شده است که امروزه از این روش برای هرس کردن درخت تصمیم استفاده میشود. از میان روشهای پس هرس درخت تصمیم نیز دو روش هرس هزینه-پیچیدگی (CCP) و هرس مبتنی بر خطا (EBP) در مقایسه با سایر روشها کارایی بیشتری دارند[2] .

هرس هزینه ـ پیچیدگی (Cost Complexity Pruning) : Breiman یک روش دو مرحله ای را پیشنهاد کرده است که ابتدا درخت تولید شده به مقدار های متفاوت هرس می شود، سپس بهترین درخت به کمک اندازه گیری دقت درختهای مختلف روی یک مجموعه داده جداگانه انتخاب میشود. دراین روش هم میزان خطا و هم مقدار پیچیدگی درخت (اندازه درخت) مدنظر است.

روش کار به این شکل است که هر نود در درخت، نقطه آغاز زیر درخت های دیگر است که به برگها ختم میشوند. قبل از هرس همه نمونه های برگها متعلق به یک کلاس است اما بعد از هرس برگها ممکن است شامل نمونه هایی از کلاسهای مختلف باشند که در اینصورت کلاس غالب به عنوان برچسب کلاس برگ انتخاب می شود. میزان خطای یک برگ به تناسب تعداد نمونه های مجموعه آموزشی که متعلق به آن کلاس نمی باشند محاسبه میشود. ا گر زیر درختی هرس شود خطای مورد انتظار خطای همان نود آغازین است که بدین ترتیب تبدیل به برگ میشود واگر زیر درخت هرس نشود میزان خطا برابر میانگین خطا در برگها با وزن تعداد نمونه ها در هر برگ می‌باشد. همیشه درمجموعه آموزشی هرس باعث افزایش میزان خطا می شود و این افزایش معیاری برای اهمیت زیردرخت می باشد. در یک زیر درخت نسبت افزایش خطا به تعداد برگهای آن معیاری از کاهش خطای هر برگ بدست می دهد، که معیار هزینه-پیچیدگی می باشد.

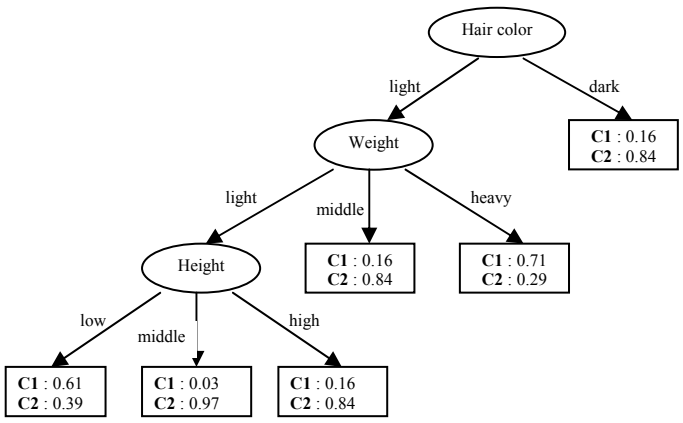
هرس مبتنی بر خطا (Error Based Pruning) : متد هرس EBP در ماشین C4.5 مبتنی بر نسبت بهره بکار گرفته شده است . این روش از اطلاعات مجموعه آموزشی برای ساخت و ساده سازی درخت استفاده می کنند. EBP نودهای درخت T را با استراتژی پائین به بالا و (Post–order) ملاقات میکند. یک ویژگی جدید EBP اینست که میتواند علاوه بر هرس معمولی، ساده سازی درخت T را با جایگزینی یک شاخه از زیردرخت T_t به جای نود پدر یعنی خود t انجام دهد(grafting) . از مزایای این متد در مقایسه با دیگر روشها اینست که امکان جایگزین کردن یک زیردرخت را با یکی از شاخه هایش به ما میدهد. این روش ، نودهای موجود در میانه درخت که مفید نبوده و ممکن است در اثر نویز در بدنه درخت وارد شده باشد را حذف می کند.

درخت تصمیم فازی

راه حل دوم برای مقابله با نویز و ناکامل بودن مجموعه آموزشی توسعه درخت تصمیم فازی است. یکی از چالش های درخت تصمیم فازی تولید پایگاه قوانین فازی بصورت اتوماتیک بخصوص در مورد یادگیری از نمونه ها می‌باشد. در درخت تصمیم فازی هیوریستیک توسعه درخت مبتنی بر ترکیب تئوری اطلاعات یا آمار با منطق فازی می باشد که در ادامه دو مدل آن بررسی میشود.

الگوریتم فازی ID3 (FID3) : این الگوریتم توسعه یافته ID3 می باشد که برروی مجموعه‌های فازی (مجموعه‌های با درجه عضویت) اعمال می شود و یک درخت تصمیم فازی را تولید می کند [4,5]. یک درخت تصمیم فازی متشکل از گره‌ها

بعنوان خصیصه آزمون ، لبه ها برای انشعاب ناشی از مقادیر فازی (ارزشهای فازی خصیصه آزمون) و برگها بعنوان برچسب کلاس ها با مقدار قطعیت می باشد. یک مثال از درخت تصمیم فازی در شکل ۱ نشان داده شده است.



شکل ۱- درخت تصمیم فازی

الگوریتم توسعه درخت تصمیم فازی بسیار شبیه ID3 می باشد با این تفاوت که ID3 خصیصه ها را براساس بهره اطلاعاتی انتخاب می کند که با احتمال معمولی داده ها محاسبه می شود، ولی در الگوریتم فازی این مقدار به کمک احتمال مقادیر عضویت داده ها محاسبه میشود. فرض کنید که مجموعه داده‌ها D باشدکه در آن هر داده دارای L خصیصه عددی پیوسته $A_1,A_2,\dots A_L$ ، یک کلاس از مجموعه $C=\{C_1, C_2, \dots Cn\}$ و مجموعه‌های فازی $F_{i1}, F_{i2}, \dots F_{im}$ برای هر خصیصه Ai (که m درهر خصیصه متغیر است) می باشد. در اینصورت اگر D^{ck} یک زیر مجموعه فازی D با کلاس C_k، و |D| مجموع مقادیر عضویت درمجموعه فازی باشد دراینصورت یک الگوریتم برای تولید درخت فازی بصورت زیر است:

۱) تولید ریشه درخت که مجموعه ای از همه داده می باشد.(یعنی یک مجموعه فازی از همه داده ها با درجه عضویت یک)

۲) اگر گره t با یک مجموعه فازی از داده D یکی از شرایط زیر را برآورده کند، بعنوان یک برگ شناخته شده و بایستی برچسب کلاس به آن نسبت داده می‌شود.

۲-۱) مقادیر عضویت داده‌های کلاس C_k نسبت به کل نمونه‌ها بزرگتر یا مساوی مقدار آستانه θ_r باشد یا $\frac{|D^{C_K}|}{|D|} \geq \theta_r$.

۲-۲) مجموع مقادیر عضویت کل داده‌ها کمتر از مقدار آستانه θ_n باشد یعنی $|D| < \theta_n$.

۲-۳) خصیصه دیگری برای افزایش دقت کلاسبندی وجود نداشته باشد .

۳) اگر سه شرط فوق برآورده نشود ، گره مورد نظر یک گره آزمون می باشد که توسعه درخت بصورت زیر دنبال میشود.

۳-۱) برای Ai ها ($i = 1,2,\dots,l$) بهره اطلاعاتی $G(A_i, D)$ بروش زیر محاسبه شده و خصیصه آزمون A_{max} که این مقدار را بیشینه می کند انتخاب میشود .

۳-۲) D را با توجه به A_{max} به زیر مجموعه های فازی $D_1, D_2, \dots D_m$ تقسیم می کنیم و درجه عضویت داده های Dj عبارتست از حاصلضرب درجه عضویت داده های D در تابع عضویت $F_{max,j}$ ، که درجه عضویت ارزشهای مختلف A_{max} در D می باشد.

۳-۳) گره‌های جدید $t_1,t_2,\dots t_m$ برای زیر مجموعه های فازی $D_1, D_2, \dots, D_m$ تولید شده و مجموعه فازی $F_{max,j}$ به لبه های بین گره‌های t و t_j انتساب داده می‌شود.

۳-۴) مجموعه D را با مجموعه D_j ($j = 1,2,\dots,m$) جایگزین کرده از مرحله ۲ بصورت بازگشتی مراحل تکرار میشود.

بهره اطلاعاتی $G(A_i,D)$ برای خصیصه Ai در مجموعه فازی D بصورت زیر تعریف میشود

$$G(A_i,D) = I(D) - E(A_i,D)$$
$$I(D) = -\sum_{k=1}^{n}(P_k.\log_2 P_k)$$
$$E(A_i,D) = \sum_{j=1}^{m}(P_{ij}.I(D_{F_{ij}})) \qquad (۱)$$
$$P_k = \frac{|D^{C_K}|}{|D|}, P_{ij} = \frac{|D_{F_{ij}}|}{\sum_{j=1}^{m}|D_{F_{ij}}|}$$

برای نسبت دادن برچسب کلاس به برگ متدهای مختلفی وجود دارد که چند روش آن بصورت زیر است :
الف) یک گره بوسیله کلاسی برچسب زده میشود که بیشترین درجه عضویت را در آن گره داشته باشد، و از بقیه داده های گره جاری صرفنظر میشود . ب) اگر شرط (۱) در مرحله دوم الگوریتم فوق برآورده شود روشی شبیه الف را در پیش می گیریم درغیر اینصورت این گره بعنوان یک گره خالی تلقی شده و از داده های آن صرفنظر میشود ج) گره بوسیله همه کلاس ها با درجه عضویت های آنها برچسب زده شده و دراینصورت همه داده های گره جاری لحاظ میشوند.

هیوریستیک Yuan برای ساخت درخت تصمیم فازی : یکی دیگر از هیوریستیک های موجود برای ساختن درخت تصمیم فازی بوسیله Yuan پیشنهاد شده است [4] . این متد بجای استفاده از کاهش انتروپی فازی ، از کمترین ابهام کلاسبندی[1] برای انتخاب خصیصه جهت توسعه درخت استفاده می کند. روش کار به طور خلاصه به صورت است :
فرض کنید گره غیر برگ S با n خصیصه $A^{(1)},\dots,A^{(n)}$ انتخاب شده است. هر خصیصه $A^{(K)}$ ، $(1<k<n)$ دارای m_k مجموعه فازی $A_1^{(k)}, A_2^{(k)},\dots, A_{m_k}^{(k)}$ می باشد. برای سادگی فرض کنید تنها دو زیر مجموعه فازی N وP به منظور کلاسبندی نمونهها داریم. برای ارزش هر خصیصه (مجموعه فازی) $A_i^{(k)}$ ، $(1 \le i \le m_k \ , \ 1 \le k \le n)$ فراوانی نسبیP وN در گره S بترتیب بصورت زیر تعریف میشود

$$P_i^{(k)} = M(A_i^{(k)} \bigcap P \bigcap S) / M(A_i^{(k)} \bigcap S)$$
و (۲)
$$q_i^{(k)} = M(A_i^{(k)} \bigcap N \bigcap S) / M(A_i^{(k)} \bigcap S)$$

که اگر X مجموعه جهانی ، F(X) مجموعه همه زیرمجموعه‌های فازی تعریف شده بر روی X و $A \in F(X)$ باشد در اینصورت M(A) به صورت زیر تعریف میشود

$$M(A) = \sum_{x \in X} A(x) \qquad (۳)$$

در گره مورد نظر S ، ابهام کلاسبندی $A_i^{(k)}$، $(1 \le i \le m_k \ , \ 1 \le k \le n)$ به صورت زیر تعریف میشود

$$Ambig_i^{(k)} = \min(p_i^{(k)}, q_i^{(k)}) / \max(p_i^{(k)}, q_i^{(k)}) \qquad (۴)$$

و متوسط ابهام کلاسبندی خصیصه k ام عبارتست از :

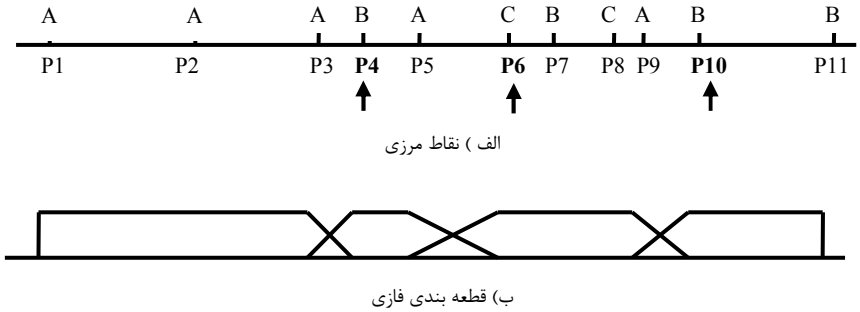
$$G_k = \sum_{i=1}^{m_k}\left(M(A_i^{(k)}) \Big/ \sum_{j=1}^{m_k} M(A_j^{(k)}) \right) Ambig_i^{(k)} \qquad k = 1,2,\dots,n \qquad (۵)$$

این هیوریستیک خصیصه ای را انتخاب می کند که متوسط ابهام کلاسبندی کمتری داشته باشد.

قطعه بندی خصیصه های پیوسته : برای جلوگیری از رشد انفجاری درخت تصمیم، رویه ساخت درخت تصمیم بایستی قابلیت تقسیم فضای خصیصه ارزش پیوسته را به تعداد دلخواه از قطعات را داشته باشد. فضای خصیصه ارزش عددی را

1- minimum classification ambiguity

می‌توان با استفاده از توابع عضویت ذوزنقه ای قطعه بندی کرد . بنابراین لازم نیست که قبل از ساختن درخت تصمیم، مقادیر فازی نمونه ها مشخص شده باشد .



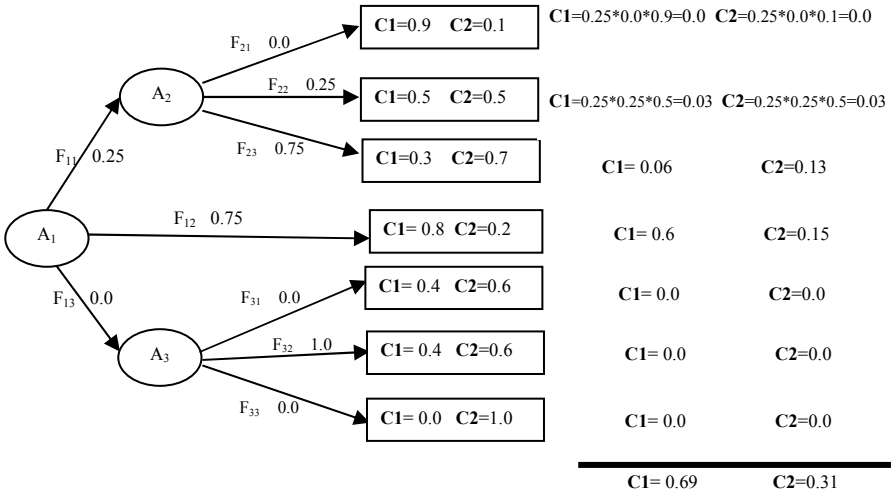
شکل۲- قطعه بندی خصیصه های با ارزش عددی پیوسته

برای افراز فضای خصیصه ارزش با مقادیر عددی پیوسته، از تابع عضویت ذوزنقـــه ای استفاده می‌کنیم. ترکیبات نامحدودی از قطعه‌های فازی برای مقادیر یک خصیصه پیوسته را می توان تولید کرد. منطقی است که نقاط مرزی را بنحوی پیدا کنیم که متعلق به کلاسهای مختلف باشند. شکل ۲-الف نقاط مرزی استخراج شده را نشان میدهد . هر نقطه مرزی بوسیله کلاس مربوطه برچسب زده شده است. حد قطعه‌بندی با استفاده از این نقاط مرزی مشخص می‌شود. نقاط مرزی مثل P3,P2 که متعلق به یک کلاس می باشند برای جستجو جهت یافتن نقاط قطعه بندی مطلوب نمی باشند بنابراین نقاطی شبیه P2 که میان دنباله ای از نقاط متعلق به کلاس یکسان قرار دارند حذف میگردد. بعنوان مثال برای تقسیم مقادیر یک خصیصه پیوسته به ۴ متغیر فازی ، ۳ نقطه مرزی بصورت تصادفی مانند شکل۲-الف (که بوسیله فلش های رو به بالا نشان داده شده است) انتخاب می شود. شکل ۲-ب متغیرهای فازی که بوسیله انتخاب این نقاط تولید شده است را نشان میدهد. پارامترها برای اعداد فازی ذوزنقه ای Trap (a,b,c,d) در این شکل به صورت زیر مشخص میشوند: متغیر زبانی دوم بوسیله نقاط مرزی P4 , P6 تعریف شده است و پارامتر a نقطه مرزی P3 و پارامتر b , P4 می باشد. پارامتر C نقطه P5 و پارامتر d نقطه مرزی P6می‌باشد. بنابراین متغیر زبانی دوم بصورت Trap(P3,P4,P5,P6) تعریف میشود. در زیر نحوه قطعه بندی خصیصه های با ارزش پیوسته با یک شبه کد ارائه شده است.

PROCEDURE for Numeric Attributes Value Space Partitioning
BEGIN
Find boundary points of the attribute value
Sort the boundary points.
For the subsequences that contain more than 2 points all belonging to the same class, remove its inner boundary points from the sorted boundary point list.
Generate all possible partitioning Combinations by selecting b-1 points from the boundary point list.
Construct b trapezoidal fuzzy numbers for each partitioning combination.
Evaluate the information gain for each trapezoidal fuzzy number set.
Return the trapezoidal fuzzy number set with the largest information gain.
END.

استنتاج تقریبی درخت تصمیم فازی : استنتاج در درخت تصمیم معمولی از گره ریشه شروع شده و با ارزیابی خصیصه‌ها در هر گره و مقدار مربوط به لبه ها ، مسیرها تا رسیدن به برگ دنبال می‌شوند و کلاس نسبت داده شده به برگ مشخص کننده کلاس نمونه مورد نظر می باشد [6]. برای تعیین کلاس یک نمونه در درخت تصمیم معمولی، تنها یک مسیر دنبال می‌شود و این درحالی است که در درخت تصمیم فازی هر گره را با بیش از یک انشعاب و ضریب قطعیت دنبال می‌شود. یک مثال در شکل ۳ نشان داده شده است که A_i خصیصه مورد ارزیابی را نشان میدهد. F_{ij} بروی لبه ها مشخص کننده مجموعه فازی و مقدار عددی بروی لبه ها نشان دهنده درجه عضویت ارزش خصیصه در مجموعه فازی می باشد. از میان

روشهای مختلف S-Norm ، برای بدست آوردن درجه عضویت یک مسیر از لبه ها و نیز برای محاسبه مقدارکلی درجه عضویت یک مسیر از لبه ها و قطعیت کلاس‌های مختلف در برگها، از ضرب استفاده می‌شود.



شکل ۳- استنتاج فازی درخت تصمیم فازی

سرانجام، برای محاسبه قطعیت کلاس مشابه از مسیرهای متفاوت، از میان روشهای مختلف T-Norm ، جمع انتخاب می‌شود. موقعی که مجموع مقدار قطعیت از یک بیشتر باشد میتوان آنها را نرمالیزه کرد. در شکل ۳ مقدار قطعیت کلی برای کلاس C_1 معادل 0.69 و برای کلاس C_2 این مقدار معادل 0.31 میباشد ، بنابراین اگر بخواهیم نمونه را از روی مقادیر فازی کلاسبندی کنیم، این نمونه متعلق به کلاس C_1 خواهد بود.

مجموعه های آموزشی

در این مقاله از مجموعه های آموزشی دانشگاه ایروین کالیفرنیا UCI [7] استفاده شده است که به شرح زیر است:

- دسته‌بندی گلهای زنبق(Iris): این مجموعه شامل ۱۵۰ نمونه از سه مدل (کلاس) مختلف از گل زنبق می باشد که تعداد نمونه ها در هر کلاس برابر ۵۰ عدد می باشد. در این مجموع چهار خصیصه با مقدار صحیح مثل طول وعرض گلبرگ بوده و با استفاده از آنها بایستی نمونه آتی در یکی از این سه کلاس ، کلاسبندی شود. در این مجموعه مقدار کمی نویز در خصیصه‌ها و کلاس نمونه ها وجود دارد.
- داده های شناسایی عینک (Glass): این پایگاه داده مشتمل بر ۲۱۴ نمونه هر یک با ۱۰ خصیصه با مقادیر پیوسته و یک برچسب کلاس توصیف شده اند . اولین خصیصه یعنی ID استفاده نمی‌شود.
- بازشناسی اعداد هفت قطعه ای 7 segment (Led) : این یک مجموعه داده مصنوعی است که بوسیله Breiman ارائه شده است. هر رقم لاتین بوسیله ۷ قطعه (دیود) نشان داده میشود که میتوانند روشن یا خاموش باشد. در این مجموعه آموزشی ده کلاس (یک کلاس برای هر رقم بین ۰ تا ۹) و هفت خصیصه دودویی (برای هر خط با مقادیر روشن و خاموش۰و۱) وجود دارد. احتمال خطای ناشی از بدکارکردن هر کدام از این هفت قطعه برابر ۱۰٪ می باشد بنابراین شانس درست بودن یک نمونه کامل ۰/۴۸=$۰/۹^{۷}$ می باشد. تعداد دلخواهی از این نمونه‌ها را میتوان با استفاده از یک برنامه تولید کرد .
- سرطان سینه (Breast Cancer) : این مجموعه داده متشکل از ۲۸۶ نمونه می باشد و مربوط به عود سرطان پستان می‌باشد . دو کلاس (عود کردن و عدم عود) و نه خصیصه داریم چهار تا از آنها عدد صحیح هستند. این خصیصه ها شامل سن ، اندازه تومور ، تعداد گره ها ، بدخیم (بلی ، خیر) ، سن یائسگی (۶۰ < ،۶۰=> ، اتفاق نیفتاده) ، پستان (

چپ ، راست) ، نتیجه اشعه (بلی ، خیر) ناحیه سینه (چپ ، راست ، بالا ، پائین ، مرکز) می باشند. در این داده ها خصیصه های گزارش نشده و نویز وجود دارد و هدف پیش بینی عود سرطان سینه بیمار بعدی است.

نتایج شبیه سازی

در آزمایشات، ابتدا مجموعه آموزشی به دو قسمت، ۱۰٪ مجموعه آزمون و ۹۰٪ مجموعه آموزشی تقسیم می‌شود (اعتبار متقاطع ۱۰ تایی) سپس درخت تصمیم بر روی مجموعه آموزشی ساخته شده و بوسیله مجموعه آزمون ارزیابی میشود. جدول ۱ نتیجه آزمایشات را با سطح اطمینان ۰/۱ نشان میدهد (لازم به ذکر است که درختهای مربوط به ستونهای اول و دوم با متد .G.R تولید شده است [1]) .

جدول ۱- میزان خطای اعمال روشهای مختلف برروی مجموعه های آموزشی (به درصد)

Method / Database	CCP	EBP	FID3	Yuan
Iris	۵/۳	۵/۳	۲/۸	۵
Glass	۳۷/۲	۳۵	۳۶/۲	۳۳/۱
Led-1000	۲۶/۲	۲۷	۲۸/۲	۲۹/۱
Led-200	۳۵	۳۲/۷	۳۴/۹	۳۶
Cancer	۲۴/۱	۲۳/۲	۲۵	۲۲/۹

در مجموعه های آموزشی Iris با توجه به موجود بودن مجموعه آموزشی فازی و نیز ابهام پائین این داده ها ، FID3 نسبت به سایر روشها دقت بیشتری دارد. در مجموعه های آموزشی Glass و Cancer مجموعه های فازی به کمک قطعه بندی (Segmentation) ایجاد شده اند (برای هر کدام از خصیصه‌ها سه متغیر فازی تعریف شده است). با توجه به ابهام بالا و وجود داده‌های گزارش نشده در این دو مجموعه آموزشی، روش Yuan دقت بیشتری نسبت به سایر روشها دارد. در مجموعه آموزشی Led با توجه به مصنوعی بودن داده ها، گسسته بودن خصیصه ها و بالا بودن مبنای خطا نسبت به تعداد نمونه ها ، درخت‌های حاصل رشد زیادی داشته و هرس آن باعث رشد چشمگیر دقت می شود. اما درخت های تصمیم فازی در این محیط کارایی مناسبی ندارند.

علاوه بر این آزمایشات ما نشان داد که داده های مصنوعی نسبت به اعمال روش هرس حساسیت بیشتری دارند و در مجموعه آموزشی Led-200 دقت روش هرس مبتنی بر خطا بیشتر است در حالیکه در مجموعه آموزشی Led-1000 دقت هرس هزینه-پیچیدگی (بدلیل ماهیت این نوع هرس) بیشتر می‌باشد.

نتیجه‌گیری

در این مقاله هرس کردن و منطق فازی بعنوان دو روش بهبود کارایی درخت تصمیم مورد بحث و بررسی قرار گرفت. درخت تصمیم فازی با استفاده از متغیر های فازی سعی می کند ضعف نمایش دانش را برطرف کند در حالیکه در هرس کردن این کار با استفاده از هیوریستیک های کاهش عمق انجام میشود. از مزایای درخت تصمیم فازی نسبت به هرس، قابلیت تقسیم نرم و از معایب آن نیاز به داشتن توابع عضویت و متغیرهای فازی است. بطور کلی در صورت داشتن توابع عضویت یا استفاده از مدل قطعه بندی در خصیصه های پیوسته، درخت تصمیم فازی دقت بیشتری از درخت های تصمیم هرس شده بوسیله هر

کدام از مدلهای معروف هرس دارد، درحالیکه اگر توابع عضویت را به صورت تصادفی انتخاب کنیم درخت تصمیم هرس شده بهتر خواهد بود. هیوریستیک های مختلف ساخت درخت فازی نیز هر کدام در مجموعه های آموزشی خاص کارایی بهتری از خود نشان میدهند بعنوان مثال FID3 بخاطر مقدار محاسبات کم درمجموعه های آموزشی بزرگ مورد استفاده قرار می‌گیرند درحالیکه هیوریستیک Yuan برای نمونه های با عدم قطعیت بالا استفاده میشود. بطور کلی نمی توان گفت که کدام هیوریستیک بهترین است و بایستی برای هر مسئله خاص هیوریستیک مناسب را انتخاب کنیم.

مراجع

[۱] ستار هاشمی، محمدرضا کنگاوری . مقایسه عملی روشهای مختلف یادگیری درخت تصمیم. دهمین کنفرانس سراسری برق ایران، تبریز اردیبهشت ماه ۱۳۸۱ ، ص ۴۲۶-۴۱۷.

[۲] ستار هاشمی، محمدرضا کنگاوری . مقایسه روشهای مختلف هرس درخت تصمیم. پنجمین کنفرانس سراسری کامپیوتر مشهد، آذر ماه ۱۳۸۱ .

[۳] ستار هاشمی، محمدرضا کنگاوری .یادگیری موازی درخت تصمیم . نهمین کنفرانس سالانه انجمن کامپیوتر ایران، تهران، بهمن ماه ۱۳۸۲ ، ص ۲۰۵-۱۹۸ .

[4] Cezary Z. Janikow. Fuzzy decision trees: Issues and methods. IEEE transaction on sytems, man, and cybernetic , vol. 28, 1998.

[5] Koen Myung lee, et al. A fuzzy decision tree induction method for fuzzy data. IEEE international fuzzy system conference proceeding Seoul, Korea, 1999.

[6] X. Z. Wang, E. C. C. Tsang, D. S. Yuang. A comparative study on heuristic algorithms for generating fuzzy decision trees. Department of computing, Hong Kong polytechnic university, 2002.

[7] UCI Repository of Machine Learning Databases.University of California, Dept. of CIS, Irvine, CA. http://www.ics.uci.edu/~mlearn/MLRepository.html .