آکادمی داده
dataacademy.ir

ایجاد پایگاه‌داده کاهش‌یافته برای داده‌کاوی فازی با استفاده از یادگیری تقویتی

محمدرضا مطش بروجردی
borujerm@aut.ac.ir
دانشگاه صنعتی امیرکبیر
دانشکده مهندسی کامپیوتر و فناوری اطلاعات

مصطفی مروج
moravvej@ce.aut.ac.ir
دانشگاه صنعتی امیرکبیر
دانشکده مهندسی کامپیوتر و فناوری اطلاعات

چکیده

در این مقاله روشی ارائه شده است که با استفاده از روش یادگیری تقویتی، مجموعه نمونه‌های اولیه آموزشی مورد استفاده در داده‌کاوی فازی را به مجموعه‌ای از الگوهای فازی با حجم بسیار کمتر تبدیل می‌کند که دارای ارزش اطلاعاتی برابر با مجموعه اولیه است. روش ارائه شده برای پایگاه‌های داده‌ای که دارای حجم بالایی از اطلاعاتی هستند، که نیاز به پردازش دارد و همچنین برای پایگاه‌های داده پویایی که دائما بر حجم اطلاعات آنها افزوده می‌شود، مناسب می‌باشد. روش ارائه شده، بمنظور برخورد با مشکلات اصلی داده کاوی همانند وجود داده‌های ناقص و رفع نویز موجود در نمونه‌های آموزشی نیز راه‌کارهایی را ارائه کرده است.

کلمات کلیدی: داده‌کاوی فازی، پیش پردازش داده‌ها، کاهش داده‌ها، یادگیری تقویتی، پایگاه داده فازی، رفع نویز

۱- مقدمه

افزایش میزان استفاده از سیستمهای کامپیوتری و سیستم‌های اطلاعاتی و بعلاوه، حجم بالای داده‌های ذخیره شده در آنها، باعث وابستگی بسیاری از فعالیت‌های تجاری به استفاده از این سیستم‌ها شده است. امروزه، مبنای بسیاری از تصمیم-گیری‌ها، اطلاعات ذخیره شده در این سیستم‌هاست، لیکن حجم بالای اطلاعات موجود در آنها، باعث شده تا پردازش و تحلیل آن برای مدیران و تصمیم‌گیران مشکل باشد. از این رو، برای کمک مدیران در پردازش این اطلاعات و دستیافتن به اطلاعات کلیدی و ارزشمند، تکنیک‌ها و ابزارهای گوناگونی بوجود آمده‌اند، که از جمله آنها می‌توان از پایگاههای داده تحلیلی و یا سیستم‌های پردازش تحلیلی آنی را نام برد، که با ارائه گزارشات کلی و نمودارها، مدلی مفیدتر از وضعیت سازمان در اختیار مدیران قرار می‌دهند. یکی از مهمترین نیازهای پردازشی، یافتن الگوهای صحیح، جدید، مفید و قابل درک از داده-هاست، بگونه‌ای که ارائه دهنده رفتارهای غیر مشهود آنها باشند. برای یافتن اینچنین الگوهایی، از تکنیک‌های داده کاوی استفاده می‌شود [۱]، [۲].

در اغلب روش‌های داده‌کاوی، پیچیده‌ترین و زمانبرترین بخش از فرایند، استنتاج مجموعه الگوهای مورد نظر از مجموعه آموزشی است. در مواردی که با مجموعه‌های حجیم اطلاعات، که نیاز به پردازش دارند، سر و کار داریم، این مساله جدی‌تر است. برای رفع این مشکل معمولا روش‌هایی مورد استفاده قرار می‌گیرند، که به نحوی، حجم داده‌های آموزشی را کاهش دهند [۳]، [۴]، [۵]، [۶]. کاهش داده‌های آموزشی به طرق مختلف مثل کاهش تعداد مشخصه‌های مرتبط با داده‌کاوی و

یا کاهش تعداد نمونه‌های آموزشی، انجام می‌شود. کاستن از مجموعه آموزشی می‌بایست به طریقی صورت پذیرد که در نتیجهٔ نهایی داده‌کاوی تاثیرگزار نباشد. بدین مفهوم که ارزش اطلاعات مجموعه کاهش یافته می‌بایست با مجموعه اولیه برابری کند [۷]. از دیگر مشکلاتی که نتایج اجرای الگوریتم‌های داده‌کاوی را تحت تاثیر قرار می‌دهد، وجود نویز و یا داده‌های ناقص[1] در مجموعه آموزشی است. برای رفع این گونه مشکلات، روش‌های رفع نویز و یا روش‌های تخمین مقادیر نامعلوم به‌کارگرفته می‌شوند[۸].

در این مقاله روشی ارائه شده است که حجم داده‌های آموزشی را برای استفاده در روش‌های داده‌کاوی فازی کاهش می‌دهد. بدین صورت که بر اساس نمونه‌های آموزشی، مجموعه‌ای از الگوهای فازی ایجاد می‌کند که از نظر حجم از مجموعه اولیه کوچکتر است، لیکن از نظر ارزش اطلاعاتی برای استفاده در روش‌های مختلف داده‌کاوی فازی، با مجموعه اولیه برابری می‌کند. در این مقاله مجموعه الگوهای نهایی حاصل از اجرای الگوریتم را با نام پایگاه‌داده کاهش‌یافته، می‌شناسیم. روش ارائه شده، یک روش افزایشی تدریجی[2] است، که در آن نمونه‌های آموزشی می‌توانند بصورت تک‌تک به مجموعه نهایی افزوده شوند. حجم پایگاه‌داده کاهش‌یافته حاصل از این روش، دارای یک حد آستانه بالا است؛ بدین مفهوم که با وجود تعداد نمونه‌های آموزشی اولیه بسیار بالا نیز حجم آن از مقدار مشخصی، افزایش نمی‌یابد. این پایگاه‌داده، مجموعه‌ای از الگوهای فازی است، افزودن یک نمونه آموزشی به آن به منزله تقویت برخی الگوها و تضعیف سایر الگوها است. تقویت یا تضعیف الگوها مبتنی بر روش یادگیری تقویتی [۹] انجام می‌گیرد. به‌منظور برخورد با عدم‌قطعیت موجود در داده‌های خام آموزشی، الگوریتم ارائه شده، بگونه‌ای است که راه‌حلی مناسب برای شناسایی و رفع نویز و همچنین تعیین مقادیر نامشخص ارائه می‌کند.

پیکر بندی این مقاله بدین صورت است که بخش ۲، روش انجام فازی‌سازی بر روی داده‌های خام را تشریح می‌کند. روش ارائه شده در این بخش بگونه‌ای است که برای ایجاد پایگاه داده کاهش‌یافته مناسب است. در بخش ۳، الگوریتم پایه ایجاد پایگاه‌داده کاهش‌یافته بیان شده است. بخش ۴، حجم پایگاه‌داده کاهش‌یافته را بررسی کرده و نشان می‌دهد که در هر حالت حجم آن از مقدار مشخصی بالاتر نمی‌رود. در بخش ۵، توسعه لازم بر روی الگوریتم پایه بمنظور برخورد با داده های ناقص ارائه شده است. بخش ۶، روش برخورد با داده‌های ناقص را به الگوریتم پایه می‌افزاید. در بخش ۷، یک نمونه از اجرای الگوریتم و نتایج حاصل از آن نشان داده شده است. جمع بندی و نتیجه گیری نیز در بخش ۸ ارائه شده است.

۲- فازی سازی داده‌ها

همانند تمامی سیستمهای فازی، در این روش نیز بمنظور پردازش فازی داده‌ها، اولین گام فازی‌سازی مجموعه نمونه‌های آموزشی است. فرض می‌کنیم که اطلاعات مجموعه اولیه در جدولی مثل T قرار دارند، بدین ترتیب که هریک از ستونهای جدول مشخص کننده ویژگی خاصی از آن نمونه و هر ردیف یک نمونه را مشخص می‌کند. در این مرحله، فازی‌سازی بر روی مقادیر مشخصه‌های عددی صورت می‌پذیرد. بدین ترتیب که درجهٔ عضویت هر یک از مقادیر عددی را به ازای یکایک مقادیر فازی موجود در دامنه آن محاسبه می‌کنیم.

تعریف: فرض کنید F_i یک مشخصه از نمونه آزمایشی باشد، مجموعه مقادیری را که در فضای فازی قابل تخصیص به F_i هستند، دامنه فازی F_i می نامیم. به بیان دیگر، دامنهٔ فازی F_i مجموعهٔ الفاظی است که F_i پس از فازی‌سازی می‌تواند اتخاذ کند.

الگوریتم فازی سازی برای مشخصه‌های یک نمونه خاص مطابق شکل زیر است:

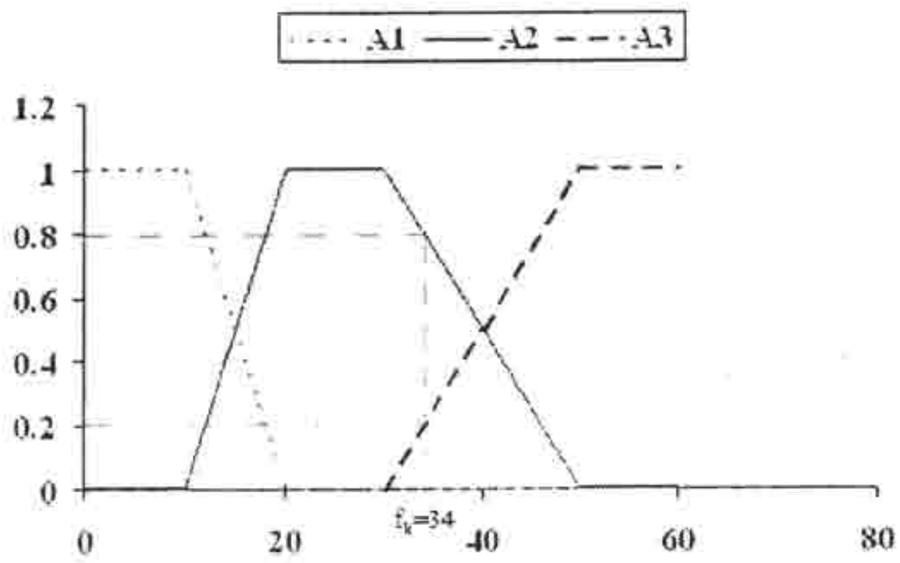
[1] Missing Values

[2] Incremental

```
For each attribute Fi in sample attribute set do
        For each fuzzy value V in fuzzy domain of Fi do
                Set membership value of fi correspond to V to μV(fi)
```

شکل ۱- الگوریتم فازی سازی نمونه‌های اولیه

که در آن μ_V تابع عضویت لفظ فازی V است. به عنوان مثال چنانچه در جدول T، مقدار یک فیلد عددی مثل F_1، بتواند مقادیر فازی A_1، A_2 و A_3 را به خود اختصاص دهد، مطابق با شکل ۲ عمل می کنیم.



T Table				
...	F_{k-1}	F_k	F_{k+1}	...
		34		

...	F_{k-1}	F_k			F_{k+1}	...
		A_1	A_2	A_3		
		0	0.8	0.2		

شکل ۲- شیوه محاسبه مقادیر عضویت فازی ارزش هر یک از مشخصه‌های نمونه‌های آموزشی به تک تک ازای الفاظ موجود در دامنهٔ آنها

به روشی مشابه به ازای تمامی نمونه‌های آموزشی میزان عضویت هر یک از مشخصه‌های جدول T را در مجموعه‌های مرتبط موجود در دامنهٔ آن، محاسبه می کنیم.

گام بعد از فازی سازی با هدف تبدیل ساختار داده‌ها به قالب مناسب برای پایگاه دادهٔ کاهش یافته صورت می‌پذیرد. در این مرحله الگوهای فازی ممکن از نمونهٔ فازی سازی شده به همراه قوت[3] هر کدام از الگوها بدین ترتیب ایجاد می‌شود که، ترکیبات مختلفی که به ازای الفاظ مختلف فازی هر یک از ستونها ممکن است بوجود آید، بروشی مشابه آنچه در شکل ۳ آمده، ایجاد می کنیم.

dataacademy.ir

جدول R_i مجموعه الگوهای بدست آمده از نمونهٔ i از مجموعه آموزشی T را نشان می‌دهد. ستون Strength در جدول R_i مشخص کنندهٔ میزان قوت هریک از الگوهای بدست آمده است. این قوت برای الگویی مثل P_k، براساس میزان عضویت هر یک از مقادیر مشخصه‌های نمونهٔ i از جدول T در مجموعهٔ فازی مقدار ارائه شده برای آن ستون در الگوی P_k و براساس فرمول زیر قابل محاسبه است:

$$Strength(P_j)=\prod_{j=1}^{L}\mu_{S_{F_j}}(f_j) \qquad (1)$$

که در آن S_{F_j} مقدار فازی ارائه شده برای ستون F_j در الگوی P_k است و f_j مقدار ستون F_j در جدول T است. در جدول R_i داریم:

[3] -Strength

$$\sum_{P_j \in R_i} Strength(P_j) = 1 \qquad (2)$$

سپس مجموعه الگوهایی از جدول R_i را که دارای قوت صفر هستند، حذف می کنیم.

بدین ترتیب یک نمونهٔ آموزشی به مجموعه‌ای از الگوهای فازی تبدیل می‌شود که هریک با قوت خاصی معتبر هستند. به شرحی که در بخش ۳ ارائه شده است، از این مجموعه الگوها برای ایجاد پایگاه داده کاهش یافته استفاده خواهد شد.

F_1			F_2			F_3		
A_1	A_2	A_3	B_1	B_2	B_3	C_1	C_2	C_3
1	.5	.4	.2	.8	0	.55	.45	0

R_i

F_1	F_2	F_3	Strength
A_1	B_1	C_1	0.011
A_1	B_1	C_2	0.009
A_1	B_1	C_3	0
A_1	B_2	C_1	0.044
A_1	B_2	C_2	0.036
A_1	B_2	C_3	0
.	.	.	.
.	.	.	.
.	.	.	.

شکل ۳ - ایجاد مجموعه الگوهای فازی

۳- پایگاه‌داده کاهش‌یافته

روش ایجاد پایگاه داده یک روش افزایشی تدریجی است که در آن نمونه‌های آموزشی جدول T می‌توانند یک به یک به پایگاه داده افزوده شود. الگوریتم کلی کار از دو گام اصلی زیر تشکیل شده است :

گام ۱- یک نمونه آموزشی مثل i از جدول T به روشی که در بخش ۲ ارائه شد، فازی سازی شده و مجموعه الگوهای فازی R_i متناظر با آن ایجاد می‌شود.

گام ۲- بر اساس مجموعه الگوهای ایجاد شده پایگاه داده کاهش یافته با شرحی که در این بخش ارائه می‌شود، ایجاد می‌شود.

ساختار پایگاه داده کاهش یافته مشابه با ساختار جدول R_i است، با این تفاوت که بجای ستون strength، ستونی به نام probability در آن وجود دارد که توسط الگوریتم مقدار دهی شده و به ازای هر الگوی فازی، احتمال رخداد آن را در مجموعه نمونه‌های آموزشی T، مشخص می‌کند. مقدار احتمال (probability) یک الگو، با استفاده از مجموعه الگوهای ایجاد شده برای هر یک از نمونه‌های آموزشی و بر اساس روش یادگیری تقویتی، محاسبه می‌شود. شکل ۴، الگوریتم ایجاد پایگاه داده کاهش یافته را برای یک نمونه آموزشی ارائه می‌دهد.

یک به یک نمونه‌های آموزشی بدین ترتیب به پایگاه داده کاهش یافته افزوده می‌شوند. مقدار probability برای یک الگوی خاص در پایگاه‌داده کاهش‌یافته به شرحی که در ادامه خواهد آمد، عددی در بازه [0, 1] است که احتمال رخداد آن الگو را در در مجموعهٔ آموزشی T بیان می‌کند. در اولین گام از الگوریتم، الگوهایی که در پایگاه‌داده کاهش‌یافته وجود ندارند، با احتمال صفر به آن افزوده می‌شوند. در گام بعدی الگوهایی از پایگاه داده که در مجموعه الگوهای R_i، منتج از فازی‌سازی ردیف i از مجموعه آموزشی وجود دارند، در پایگاه‌داده کاهش‌یافته تقویت می‌شوند. میزان تقویت هر یک از این الگوها برابر با قوت آن الگو در مجموعهٔ الگوهای موجود در R_i تقسیم بر تعداد نمونه‌های پایگاه‌داده کاهش‌یافته است. از آنجایی که مطابق (۲) مجموع قوت الگوهای R_i برابر با یک است، داریم:

۱) تمامی الگوهایی در R_i که الگوی معادل آن در پایگاه داده کاهش یافتهٔ R وجود نداشته باشد، به پایگاه داده افزوده شود و احتمال آنها برابر با صفر قرار گیرد.

۲) اگر N برابر با تعداد نمونه‌های آموزشی افزوده شده به پایگاه داده باشد آنگاه N=N+1 .

۳) میزان تقویت احتمال الگوهایی (مثل P_j) از پایگاه داده کاهش یافته R که الگوی متناظر با آن در R_i وجود دارد، برابر است با:

$$ReInf(P_j) = \frac{Strength_i(P_j) - probability(P_j)}{N.\sum_{P_k \in R_i} Strength_i(P_k)}$$

که در آن $ReInf(P_j)$ میزان تقویت احتمال الگوی P_j را در پایگاه دادهٔ نهایی مشخص می کند، $Strength_i(P_j)$ قوت الگوی P_j در جدول R_i است. احتمال رخ داد الگوی P_j مطابق فرمول زیر محاسبه می‌شود:

$$probability(P_j) = \min\left(probability(P_j) + Reinf(P_j), 1\right)$$

۴) سایر الگوهای پایگاه دادهٔ کاهش یافته R، که الگوی متناظر با آنها در R_i وجود ندارد، تضعیف می‌شوند. در این حالت برای الگویی مثل P_j، از R که معادل آن در R_i وجود ندارد، میزان تضعیف(تقویت منفی) برابر است با:

$$ReInf(P_j) = -\frac{probability(P_j)}{N}$$

احتمال رخ داد الگوی P_j مطابق فرمول زیر محاسبه می‌شود:

$$probability(P_j) = \max\left(probability(P_j) + ReInf(P_j), 0\right)$$

شکل ۴- الگوریتم ایجاد پایگاه دادهٔ کاهش یافته با استفاده از مجموعه الگوهای فازی ایجاد شده برای یک نمونه آموزشی خاص

$$\sum_{P_j \in R_i} ReInf(P_j) = \frac{1 - \sum_{P_j \in R_i} probability(P_j)}{N} \qquad (3)$$

افزودن مجموعه R_i همانطور همچنین باعث تضعیف سایر الگوهای پایگاه داده که در R_i وجود ندارند خواهد شد، در این حالت داریم:

$$\sum_{P_j \notin R_i} ReInf(P_j) = -\frac{\sum_{P_j \notin R_i} probability(P_j)}{N} \qquad (4)$$

برای اولین الگوی افزوده شده به پایگاه داده R مطابق الگوریتم فوق داریم:

$$\sum_{P_j \in R_1} ReInf(P_j) = 1, \quad \sum_{P_j \in R_1} probability(P_j) = 1 \qquad (5)$$

برای الگوی بعدی اضافه شونده به R، از آنجایی که مجموعه الگوهایی از R که عضو R_2 هستند و مجموعه الگوهایی که عضو R_2 نیستند دو مجموعه مجزا میباشند، داریم:

$$\sum_{P_j \in R_2} probability(P_j) + \sum_{P_j \notin R_2} probability(P_j) = 1 \qquad (6)$$

بنابراین با توجه به فرمولهای (۳) و (۴) می‌بینم که:

$$\sum_{P_j \in R_2} ReInf(P_j) + \sum_{P_j \notin R_2} ReInf(P_j) = 0 \qquad (7)$$

به همین ترتیب برای سایر نمونه‌های آموزشی نیز فرمول (۷) صادق است، پس در حالت کلی خواهیم داشت:

$$\sum_{P_j \in R} probability(P_j) = 1 \qquad (8)$$

بنابراین، probability هر الگو، احتمال وقوع آن الگوی خاص را در مجموعه اولیه مشخص می‌سازد. الگوریتم ارائه شده تا اینجا بگونه‌ای است که در حالتی قابل استفاده است و جواب قابل قبول می‌دهد، که داده‌های ناقص و ارزش نامشخص نداشته باشیم. الگوریتم فوق را می‌توان بگونه‌ای بهبود بخشید که برای برخورد با داده‌های ناقص مناسب باشد.

۴- حجم پایگاه داده کاهش یافته

همانطور که گفته شد، پایگاه داده کاهش یافته نهایی، مجموعه‌ای از الگوهای فازی است. این الگوها، با استفاده از ترکیبات مختلف الفاظ فازی قابل اتخاذ توسط هریک از ستونها، بوجود آمده‌اند. فرض کنید جدول T دارای L مشخصه بوده و M_i کاردینالیتی دامنه فازی ستون i از جدول T باشد، حداکثر تعداد ترکیبات ممکن از الفاظ فازی تخصیص داده شده به فرمولها، برابر خواهد بود با:

$$Maximum\ Number\ of\ Patterns = \prod_{i=1}^{L} M_i \quad (9)$$

هنگام فازی سازی نمونه‌های آموزشی، از این مجموعه ترکیبات آن دسته الگوهایی مورد استفاده قرار می‌گیرند که دارای قوت بیشتر از صفر باشند. بدترین حالت برای حجم پایگاه داده کاهش یافته زمانی رخ می‌دهد که در نتیجهٔ فازی سازی نمونه های آموزشی هر یک از الگوهای ممکن، حداقل یک بار دارای قوت بیشتر از صفر باشند، در این حالت حجم پایگاه دادهٔ نهایی برابر با تعداد الگوهای ممکن از ترکیبات مختلف الفاظ فازی ستونها خواهند بود؛ یعنی داریم:

$$Maximum\ Database\ Size = \prod_{i=1}^{L} M_i \quad (10)$$

همانطور که مشخص است، هرچند پایگاه داده به صورت افزایشی ایجاد می‌گردد لیکن حجم نهایی پایگاه داده از مقدار خاصی فراتر نمی‌رود. به عنوان مثال برای مجموعه رکوردهایی که دارای ۱۰ فیلد بوده و بطور متوسط در دامنه هر فیلد آن ۵ مجموعه فازی وجود داشته باشد حجم پایگاه داده در بدترین حالت در حدود ۱۰۰۰۰۰ رکورد خواهد بود. همانطور که مشخص است، کاهش به این حجم داده‌ای فقط در مواردی مفید است که حجم پایگاه داده اولیه بسیار بزرگ باشد.

۵- داده‌های ناقص

برای بکارگیری روش فوق در حالتی که، بعضی از مشخصه‌ها دارای مقدار نامشخص هستند، ابتدا با روشی مشابه آنچه که در بخش ۲ ارائه شد، نمونهٔ i را به مجموعه ای از الگوهای فازی تبدیل می‌کنیم. بدیهی است که امکان مشخص کردن لفظ یا الفاظ فازی مرتبط با مشخصه‌ای از جدول نمونه‌آموزشی که ارزش آن مشخص نیست (C_{missed})، وجود ندارد. برای حل این مساله چنین فرض می‌کنیم که $C_{misssed}$ هر یک از مقادیر دامنه فازی خود را با احتمال یکسان بپذیرد. با این فرض، روش فازی سازی داده‌ها را برای استفاده در مواردی که ارزشهای نامشخص داشته باشیم، به‌طریقی که در شکل ۵ ارائه شده گسترش می‌دهیم.

نتیجه اعمال روش فوق برای فازی سازی این است که، قوت حاصل از مشخصه‌هایی که مقدار معین دارند، بین الفاظ مختلف دامنه فازی مشخصه C_{missed}، تسهیم می‌شود.

در حالتی که در نمونه آموزشی، تعداد مشخصه‌هایی که دارای ارزش نامشخص هستند، بیش از یک مشخصه باشد، نیز به طریق مشابه به ازای هر یک از آن مشخصه‌ها می‌توان روش فوق را بکار گرفت. سایر گامهای الگوریتم ایجاد پایگاه داده، همانند آنچه در بخش ۳ به آن اشاره شد، بدون هیچ تغییری اجرا می‌شوند.

۱- نمونه آموزشی i را با روش ارائه شده در بخش قبل، بدون در نظر گرفتن مشخصه C_{missed}، به مجموعه الگوهای فازی R_i متناظر با آن تبدیل می‌کنیم.

۲- با فرض اینکه C_{missed} هر یک از مقادیر دامنه فازی خود را می‌تواند به خود اختصاص دهد، مجموعه الگوهای R_i را با ترکیب تمامی مقادیر دامنه فازی C_{missed} و الگوهای R_i توسعه می‌دهیم.

۳- برای محاسبهٔ قوت هر یک از الگوهای موجود در R_i، از فرمول زیر استفاده می‌کنیم:

$$Strength_i(P_j)=\frac{1}{TN_{C_{missed}}}\left(\prod_{F_k\neq C_{missed}}\mu_{S_{F_k}}(f_k)\right)$$

که در آن، S_{F_k} مقدار فازی ارائه شده برای مشخصه F_k در الگوی P_j است. f_k ارزش مشخصه F_k در جدول T و $TN_{Cmissed}$ تعداد الفاظ فازی موجود در دامنه مشخصه C_{missed} است.

شکل ۵ - الگوریتم ایجاد پایگاه داده کاهش یافته در صورت وجود مقادیر ناقص

۶- رفع نویز

از آنجایی که نویز موجود در مجموعهٔ آموزشی اولیه، باعث ایجاد الگوهای نامتجانس با سایر الگوها می‌شود، در قالب الگوهای فازی حاصل از فازی‌سازی و در پایگاه داده کاهش یافته، نویز بصورت مجموعه‌ای از الگوها است که با سایرین تفاوت عمده داشته و نرخ رخ دادن آن نسبت به سایر الگوها کم است. با این شرح، رفع نویز در دو مرحله از اجرای الگوریتم قابل انجام است:

الف) قبل از افزودن هر یک از الگوها به پایگاه داده

پس از ایجاد مجموعهٔ الگوهای R_i از نمونه آموزشی i، الگوهایی را که دارای قوت کمتری نسبت به سایر الگوها باشند از R_i حذف می‌کنیم.

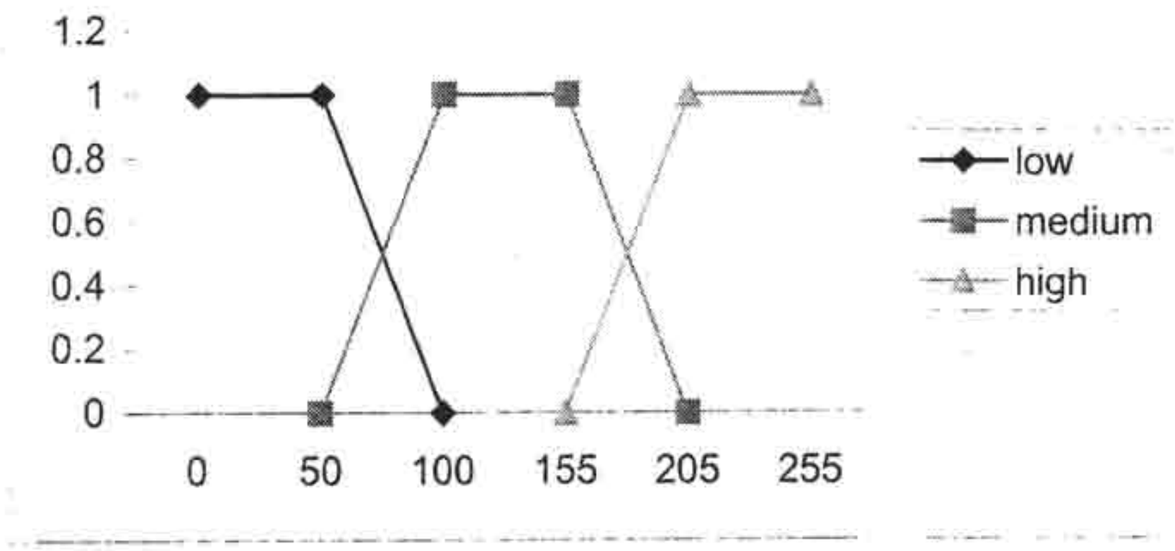
ب) پس از افزودن تمامی الگوها به پایگاه داده

پس از افزودن تمامی الگوها به پایگاه داده R، با توجه به اینکه probability هر الگو، احتمال رخ دادن آن الگو را در مجموعه آموزشی اولیه نشان می‌دهد و همانطور که قبلا بیان شد، نویز الگوهایی با احتمال وقوع پایین ایجاد می‌کند، بنابراین برای حذف نویز می‌توانیم الگوهای کم احتمال را از مجموعه R حذف کنیم.

هرچند استفاده از این روش، تعیین الگوی نویز را تسهیل می‌نماید لیکن، پیش نیاز مهم آن این است که، نمونه‌های آموزشی موجود در جدول T بطور تقریبا یکنواختی در فضای حالات توزیع شده باشند. در غیر اینصورت الگوهایی که نمونه‌های اندکی از آن در T وجود داشته باشند نیز، ممکن است به عنوان نویز شناخته شوند.

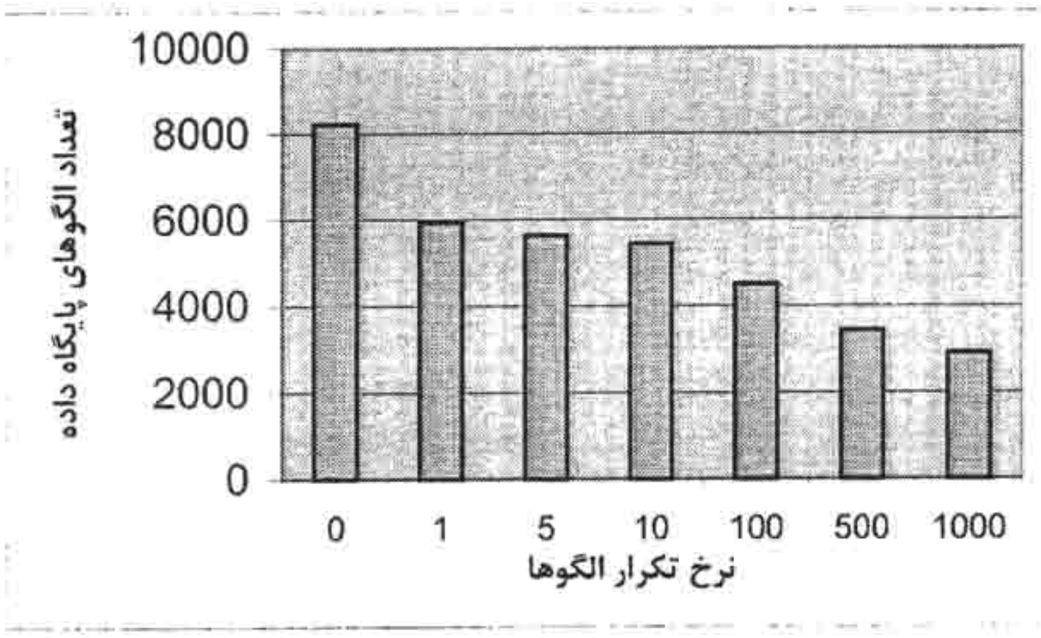
۷- نتایج اجرایی

برای آزمایش الگوریتم از یک مجموعه از نمونه‌های آموزشی، که نقاط واقع در همسایگی‌های ۳×۳ از تصویر خاکستری Lena بودند، استفاده شد. هر یک از ۶۵۰۲۲ نمونه آموزشی دارای ۹ مشخصهٔ عددی بودند که می توانستند مقادیر صفر تا ۲۵۵ را داشته باشند. در دامنهٔ فازی هر یک از مشخصه‌ها سه مقدار فازی low، medium و high با توابع عضویت ارائه شده در شکل زیر وجود داشت.



شکل ۶- توابع عضویت مجموعه‌های فازی استفاده شده برای فازی‌سازی نمونه‌های آموزشی

مطابق با الگوریتم، حداکثر حجم پایگاه دادهٔ کاهش یافته می‌بایست ۱۹۶۸۳ الگو باشد. اجرای الگوریتم فوق برای این نمونهٔ آموزشی در اولین گام و بدون حذف هیچ یک ازالگوهای ارائه شده به عنوان نویز باعث ایجاد ۸۲۱۴ الگو در پایگاه داده کاهش یافته گردید، که حدودا معادل ٪۲۵ حجم اولیه مجموعه آموزشی است. با حذف الگوهایی که فقط یک بار در مجموعه آموزشی مشاهده شده‌اند، این حجم، به ۵۹۴۲ الگو کاهش یافت که در حدود ٪۱۰ حجم اولیه مجموعه آموزشی است. شکل ۷، حجم پایگاه داده را به ازای حذف الگوهایی با نرخ‌های تکرار متفاوت نشان می‌دهد. همانطور که در شکل مشخص است، حذف الگوهایی با نرخ تکرار کم، باعث کاهش حجم پایگاه داده به میزان قابل توجهی خواهد شد.



شکل ۷- حجم پایگاه داده کاهش یافته به ازای حذف الگوهایی با نرخ‌های تکرار متفاوت

۸- نتیجه گیری

روش ارائه شده در این مقاله، از یک مجموعه از نمونه‌های آموزشی، یک پایگاه داده کاهش فازی کاهش یافته ایجاد می‌کند که در روشهای مختلف داده‌کاوی فازی قابل استفاده است و از نظر ارزش اطلاعاتی، برای استفاده در کاربردهای داده کاوی فازی، با مجموعه اولیه برابری می‌کند. مزایای الگوریتم و پایگاه داده فازی حاصل از آن عبارتند از:

۱- وجود یک حد آستانه‌ای بالا برای حجم پایگاه داده

۲- داشتن مشخصه probability، که مشخص کننده احتمال وقوع یک الگوی خاص در نمونه‌های آموزشی اولیه است.

۳- افزایشی بودن فرایند ایجاد پایگاه داده که برای مواردی که مجموعه آموزشی اولیه، افزایشی و دارای حجم بالای اطلاعات است، مناسب است.

٤- ارائه روشی برای تعیین مقادیر داده‌های ناقص

۵- ارائه روشی برای برطرف نمودن نویز موجود در داده‌ها

از آنجایی که یکی از زمانبرترین فعالیت‌ها در فرایند داده کاوی، پردازش نمونه‌های اولیه و یادگیری بر اساس آنها است، داشتن اطلاعات با حجم کم و دقت اطلاعاتی برابر با اطلاعات اولیه، ما را در سرعت بخشیدن به فرایند داده‌کاوی یاری خواهد نمود.

۹- تقدیر و تشکر

بخشی از این پژوهش، با پشتیبانی مالی مرکز تحقیقات مخابرات ایران انجام پذیرفته است که بدین‌وسیله از این مرکز محترم تشکر می‌شود.

۱۰- فهرست مراجع

[۱] Han J., Kamber M., *Data Mining concepts and techniques*, Morgan Kaufmann, 2001

[۲] Kandartzic M., *Data Mining Concepts, Models, Methods, and Algorithms*, IEEE Press, 2003

[۳] Bird R., *Fuzzy data analysis method for large volume of data*, Project report in support of degree of master of engineering, University of Bristol, Department of Engineering Mathematics, 2003

[۴] Chen J.H., Ho S.Y., *Intelligent Multi-Objective Evolutionary Algorithm for Editing Minimum Reference Set*, Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Communications of the Institute of Information and Computing Machinery, V. 5, No. 2, pp. 4-13, 2002

[۵] Hall L.O., Chawla N., Bowyer K.W., *Decision Tree Learning on Very Large Data Sets*, IEEE International Conference on Systems, Man and Cybernetics , 1998.

[۶] Wu T., *Sampling in Data Mining*, Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Communications of the Institute of Information and Computing Machinery, V. 5, No. 2, pp.95-100, 2002.

[۷] Bhm C., *Similarity Search and Data Mining: Database Techniques Supporting Next Decade's Applications*, (Keynote Speech) Proc. 4th Int. Conf. on Information Integration and Web-based Applications & Services (IIWAS), Bandung, Indonesia 2002.

[۸] Chen Z., *Data Mining and Uncertain Reasoning*, John Wiley & Sons INC, 2001

[۹] Berenji H.R., Vengerov D., *On convergence of fuzzy reinforcement learning*, Proceedings of the 10th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2001

[۱۰] Hauke K., Owoc M.L., Pondel M., *Building Data Mining Models in the Oracle 9i Environment*, Informing Science, June 2003