دسته‌بندی داده‌های جریانی فازی با استفاده از تحلیل پوششی داده‌ها

امینه توحیدی[1]، محمد امین ادیبی[2]، علیرضا علی نژاد[3,*]

[1]دانشجوی کارشناسی ارشد ، دانشگاه آزاد اسلامی، واحد قزوین ،گروه مهندسی صنایع ، قزوین،ایران
[2]استادیار دانشگاه آزاداسلامی، واحد قزوین،گروه مهندسی صنایع ، قزوین، ایران
[3]دانشیار ،دانشگاه آزاد اسلامی، واحد قزوین، گروه مهندسی صنایع، قزوین،ایران(عهده‌دار مکاتبات)
تاریخ دریافت: اسفند1394، اصلاحیه: خرداد 1395، پذیرش: مرداد 1395

چکیده

در این تحقیق یک روش دسته‌بندی داده‌های غیرقطعی از نوع فازی که از جمله چالش‌برانگیزترین حوزه‌های تحلیل داده محسوب می‌شود، ارائه شده است. در واقع حجم بالا و پیچیدگی روش‌های تحلیل داده مانع از توسعه روش‌هایی جهت تحلیل داده‌های فازی می‌شود. با این‌حال در برخی حوزه‌های دیگر همچون برنامه‌ریزی ریاضی پیشرفت‌های چشمگیری در مدل سازی سیستم‌هایی که داده‌های فازی از آنها در اختیار است، بدست آمده است و لذا توجهات به سمت بهره‌گیری از این فرصت در سال‌های اخیر جلب شده است. به منظور بهره برداری از یافته های تحقیقاتی پیرامون مدلهای ریاضی فازی، در این تحقیق یک روش جدید دسته‌بندی داده‌های فازی مبتنی بر تحلیل پوششی داده زمانی که داده‌ها به صورت جریانی وارد می‌شوند، ارائه می‌شود. روش پیشنهادی می‌تواند با استفاده از به‌هنگام‌سازی معیارهایی پیش‌بینی دسته‌ی داده‌های فازی به دسته‌بندی داده‌هایی بپردازد که در طول زمان تغییراتی در الگوی رفتاری آنها بوجود می‌آید. روش جدید توسط داده‌های شبیه‌سازی شده مورد آزمون قرار گرفته و نتایج نشان‌دهنده‌ی قابلیت این روش در مواجه با شرایط غیرقطعی و متغیر است.

کلمات کلیدی: دسته‌بندی، داده‌های فازی، تحلیل پوششی داده‌ها، برنامه ریزی ریاضی، جریان داده.

1- مقدمه

مساله دسته‌بندی[1] عبارتست از تخصیص یک کلاس یا دسته به یک داده(موقعیت) براساس یک تابع یا مدل از پیش تعیین شده. چنین مدلی از طریق مقایسه کلاس یکسری داده (که کلاس آنها مشخص می باشد یا اصطلاحا داده‌های آموزشی) از روش های مختلفی همچون درخت تصمیم، شبکه های عصبی مصنوعی، ماشین بردار پشتیبان، رگرسیون لجستیک و... حاصل می‌شود [1]. با در اختیار داشتن چنین مدلی مشخص می‌شود که یک داده به یک کلاس خاص تعلق دارد یا خیر؟ بطور خاص می توان کنترل کیفیت کامپیوتری، پایش وضعیت ماشین آلات، طراحی فرایندهای تولیدی و ... را از کاربردهای خاص دسته بندی در صنعت دانست [12].

از چالش های دسته‌بندی، ایجاد مدل های دسته‌بندی کننده برای داده‌های غیر قطعی فازی می باشد. زیرا در بسیاری از مواقع داده‌های جمع آوری شده بدلایل مختلف قطعی نیستند. از طرفی در بسیاری از مواقع داده‌هایی که باید دسته‌بندی شوند یکجا آماده نیستند بلکه درطول زمان مشاهده می‌گردند و اصطلاحا داده‌های جریانی[2] می‌باشند[7]. در چنین شرایطی سوال این است که چگونه می‌توان مدلی برای تعیین کلاس داده‌هایی که بصورت جریانی مشاهده می شوند و در ضمن دارای جنبه های غیرقطعی بصورت فازی هستند بدست آورد و آن را در طول زمان برحسب تغییرات شرایط سیستم تحت مطالعه به روز رسانی نمود؟

در چنین شرایطی استفاده از قابلیت های برنامه ریزی خطی درشرایط وجود داده‌های فازی از یکسو و قابلیت به روز رسانی سریع جواب بهینه در صورت تغییر بردارها و ماتریسهای ضرایب به کمک نرم‌افزارهای قدرتمند حل مدلهای برنامه ریزی ریاضی، دسته‌بندی براساس روش تحلیل پوششی داده‌ها (DEA) را به روشی مناسب تبدیل می کند. هرچند که روش DEA اساسا برای محاسبه کارائی نسبی تعدادی واحد هم سنخ که تعدادی نهاده را به تعدادی خروجی تبدیل می کند اما با تغییر تفسیر واحد ها می‌توان در کاربردهای دیگری از جمله دسته‌بندی داده نیز از آن استفاده کرد. در این روش هر داده به‌عنوان یک DMU در نظر گرفته می‌شود که ورودی های آن مشخصه های داده و خروجی کلاس داده است. نشان داده می‌شود فضایی که داده‌های متعلق به یک

1- Classification
*alinezhad@qiau.ac.ir

2- Streaming Data

کلاس در آن قرار می گیرند، همان فضای پوشش داده شده توسط مرز کارائی حاصل از حل یک مساله DEA است [10] با استفاده از روش DEA برای دسته‌بندی، حصول تابع دسته‌بندی کننده از حل یک مساله برنامه ریزی خطی ممکن می‌گردد. این موضوع باعث می‌گردد تا بتوان از ویژگی های برنامه ریزی خطی، از جمله تحلیل حساسیت جواب در صورت تغییر ضرایب، بصورت گسترده در دسته‌بندی استفاده کرد. از جمله این ویژگی ها بررسی داده‌های غیرقطعی فازی می باشد که در مدل های برنامه ریزی ریاضی به راحتی با آنها می‌توان برخورد نمود [9].

با بررسی ادبیات استفاده از تحلیل پوششی داده در دسته‌بندی می‌توان دریافت تحقیقات معدودی در این حوزه صورت پذیرفته است که تعدادی از آنها به معرفی این رویکرد جدید و تعدادی به استفاده از آن برای داده‌های غیرقطعی پرداخته اند. بر اساس تحقیقات انجام گرفته توسط مولفین این مقاله از مدل برنامه ریزی ریاضی مربوط به DEA در بررسی دسته‌بندی داده‌های جریانی استفاده نشده است که در این تحقیق به آن پرداخته می شود. برای این منظور یک مکانیزم جدید باید پیشنهاد گردد تا بتواند مدل برنامه ریزی ریاضی جدید را برای دسته‌بندی داده‌های جریانی بکار بگیرد. این مکانیزم باید بتواند با کنترل ورود و خروج داده‌های موثر بر مرز کارائی در طول زمان، بستر مناسب بهینه سازی را فراهم کند. در این خصوص می‌توان یک روش خلاصه سازی داده‌ها (فشرده سازی) از قبیل روش‌های خوشه بندی[3] [4] را پیشنهاد داد که از یک سو بردارهای داده ورودی را دریافت و آنها را در بردارهایی که می‌تواند مرکز خوشه ها باشد، خلاصه و برای درنظر گرفته شدن در دسته‌بندی ارسال شود [2]، [5].

در ادامه این مقاله در بخش 2 دسته بندی با DEA ارائه می شود و در بخش 3 حالت فازی بررسی می گردد. چارچوب پیشنهادی برای دسته بندی داده ای جریانی قرار داده شده است. روش پیشنهادی در بخش 5 مورد آزمون قرار می گیرد و نتایج در بخش 6 گنجانده شده است.

2- دسته‌بندی داده با استفاده از تحلیل پوششی داده

فرض کنید می خواهیم به هدف دسته‌بندی محدوده یک دسته را شناسائی کنیم. حال اگر هر داده یک DMU در نظر گرفته شود که مقادیر مشخصه های هر داده به‌عنوان ورودی DMU و عدد یک هم به‌عنوان تنها خروجی آن باشد، با حل یک مساله DEA می‌توان داده‌هایی که به‌عنوان نقاط مرزی در DEA شناخته می شوند را برای ترسیم محدوده دسته مورد استفاده قرار داد و سپس از این محدوده ها برای پیش بینی دسته داده‌های جدید استفاده کرد. یعنی اگر n داده $x_i = (x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{im})$ که $i = 1,2, \dots, n$ در یک دسته باشند می‌توان با حل مجموعه مساله های برنامه ریزی خطی نشان داده شده در معادله (1) در قالب یک مساله DEA برای تشخیص نقاط مرزی برای تعیین محدوده دسته اقدام کرد.

به‌عنوان مثال مجموعه داده‌های جدول (1) را که در دو دسته 1 و 2 هستند را در نظر بگیرید. این داده‌ها از [8] انتخاب شده اند.

برای شناسائی محدوده های دسته‌های دوگانه دو سری مساله های DEA معرفی شده در (2) و (3) را حل می کنیم که مدل‌های ورودی محور و خروجی محور هستند. محدوده ی بدست آمده برای دو دسته 1 و 2 بر اساس نقاط مرزی نیز در شکل (1) نشان داده شده است. در ضمن مقادیر متغیر وابسته θ در حل دو سری مسائل LP مربوط به دسته اول و دوم در جداول (2) و (3) آمده است که از روی آنها شناسائی نقاط مرزی ممکن می‌گردد.

$$Minimize\theta^t$$

Subject to:

$$\sum_{i=1}^{n} \lambda_i x_{ij} - \xi^t x_{tj} \le 0, \quad j = 1, \dots, m$$

$$\sum_{i=1}^{n} \lambda_i = 1$$

$$\lambda_i \ge 0\ ,\ i = 1, \dots, n. \qquad (1)$$

جدول (1): داده‌های نمرات داوطلبین ورود به دانشگاه

دسته (رد)	مشخصه دوم (نمره GPA)	مشخصه اول (نمره GMAT)	دسته (قبول)	مشخصه دوم (نمره GPA)	مشخصه اول (نمره GMAT)
2	4.7	310	1	6.02	640
2	4.5	350	1	6.09	550
2	4.7	400	1	5.67	510
2	4.8	370	1	5.54	420
2	4.7	450	1	6.75	560
2	4.5	500	1	6.6	550
2	4.6	520	1	5.87	580
2	4.3	550	1	6.2	420
2	4.5	570	1	6.77	450
2	4.9	450	1	5.67	520
2	4.6	320	1	5.33	440
2	4.6	400	1	5.96	480
2	5.1	310	1	6.13	520
			1	6.26	570
			1	5.95	400
			1	5.2	580

$$Minimize\theta^t \qquad (2)$$

3- Clustering

Subject to:

$$\sum_{i=1}^{n} \lambda_i x_{ij} - \theta^t x_{tj} \leq 0, \quad j = 1,2$$

$$\sum_{i=1}^{16} \lambda_i = 1$$

$$\lambda_i \geq 0 , i = 1, \dots ,16.$$

$$Maximize \theta^t$$

Subject to:

$$\sum_{i=1}^{n} \lambda_i x_{ij} - \theta^t x_{tj} \geq 0, \quad j = 1,2$$

$$\sum_{i=1}^{13} \lambda_i = 1$$

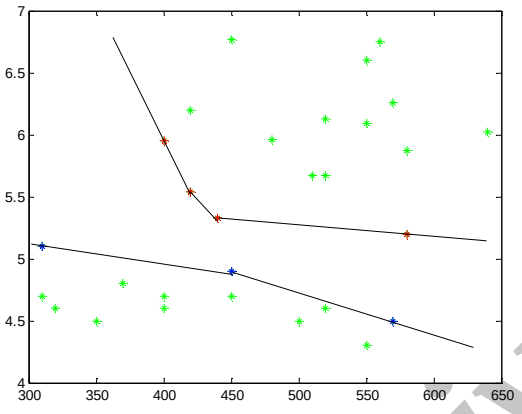
$$\lambda_i \geq 0 , i = 1, \dots ,13. \qquad (3)$$

جدول(2): مقادیر متغیر θ در حل سری مسائل LP مربوط به دسته اول

θ	شماره داده (دسته **1**)
0.87	1
0.87	2
0.93	3
1.00	4
0.79	5
0.81	6
0.90	7
0.96	8
0.89	9
0.93	10
1.00	11
0.90	12
0.87	13
0.85	14
1.00	15
1.00	16

جدول (3): مقادیر متغیر θ در حل سری مسائل LP مربوط به دسته دوم

θ	شماره داده (دسته**2**)
1.08	1
1.11	2
1.05	3
1.04	4
1.03	5
1.04	6
1.01	7
1.04	8
1.00	9
1.00	10
1.10	11
1.07	12
1.00	13



شکل (1): محدوده ی بدست آمده برای دو دسته 1 و 2 بر اساس نقاط مرزی

3-دسته‌بندی داده‌های غیرقطعی فازی با استفاده از تحلیل پوششی داده

با فرض این که مقدار مشخصه ی j ام $(j = 1,2, \dots ,m)$ داده iام $(i = 1,2, \dots ,n)$ بصورت یک عدد فازی ذوزنقه ای به فرم $\tilde{x}_{ij} = (a_{ij}, b_{ij}, c_{ij}, d_{ij})$ باشد که تابع عضویت آن در شکل (2) نشان داده شده است، آنگاه الگوی مدل برنامه ریزی خطی متناسب با یک مساله DEA به شکل معادله 4 خواهد بود.

$$minimize \ \theta^t$$

Subject to:

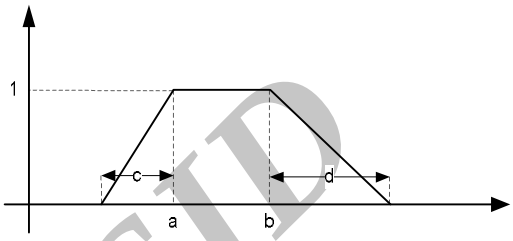
$$\sum_{i=1}^{n} \lambda_i \tilde{x}_{ji} \leq \theta^t \tilde{x}_{jt}, \quad j = 1, \dots ,m \qquad (4)$$

$$\sum_{i=1}^{n} \lambda_i \tilde{y}_{ri} \geq \tilde{y}_{rt}, \quad r = 1, \dots, s$$

$$\sum_{i=1}^{n} \lambda_i = 1$$

$$\lambda_i \geq 0, \quad i = 1, \dots, n$$

در [6] یک روش برای حل مساله نشان داده شده در معادله 4 ارائه شده است که با حل آن یک مقدار قطعی برای مقادیر θ بدست می آید. مدل مورد نظر که یک مدل LP قطعی است در 5 آمده است که در آن $x^L = a$، $x^R = b$، $\alpha^L = c$ و $\alpha^R = d$ و همچنین s تعداد خروجی های هر DMU است.



شکل(2): درجه عضویت عدد فازی ذوزنقه ای $\tilde{x} = (a, b, c, d)$

جدول (4): مقادیر 20 داده فازی در دو دسته 1و 2

شماره داده	x1_a	x1_b	x1_c	x1_d	x2_a	x2_b	x2_c	x2_d	شماره دسته
1	3.25	4.00	0.50	0.50	3.50	4.25	0.50	0.50	1
2	3.00	4.00	0.50	0.50	4.75	5.25	0.50	0.50	1
3	3.00	4.00	0.50	0.50	5.75	6.25	0.50	0.50	1
4	2.00	3.00	0.50	0.50	5.75	6.25	0.50	0.50	1
5	4.00	5.00	0.50	0.50	4.75	5.25	0.50	0.50	1
6	5.00	6.00	0.50	0.50	3.00	3.50	0.50	0.50	1
7	5.00	6.00	0.50	0.50	6.00	6.50	0.50	0.50	1
8	4.00	5.00	0.50	0.50	3.75	4.25	0.50	0.50	1
9	4.96	5.96	0.50	0.50	5.31	5.81	0.50	0.50	1
10	5.26	6.26	0.50	0.50	5.38	5.88	0.50	0.50	1
11	1.50	2.50	0.50	0.50	0.50	1.00	0.50	0.50	2
12	1.25	2.25	0.50	0.50	1.75	2.25	0.50	0.50	2
13	1.25	2.25	0.50	0.50	2.75	3.25	0.50	0.50	2
14	0.25	1.25	0.50	0.50	2.75	3.25	0.50	0.50	2
15	2.25	3.25	0.50	0.50	1.75	2.25	0.50	0.50	2
16	3.25	4.25	0.50	0.50	2.00	2.50	0.50	0.50	2
17	3.25	4.25	0.50	0.50	2.00	2.50	0.50	0.50	2
18	2.25	3.25	0.50	0.50	0.75	1.25	0.50	0.50	2
19	3.21	4.21	0.50	0.50	2.31	2.81	0.50	0.50	2
20	3.51	4.51	0.50	0.50	2.38	2.88	0.50	0.50	2

به‌عنوان مثال داده‌های جدول (4) که مربوط به 20 داده فازی ذوزنقه ای در دو دسته یک و دو هستند را در نظر بگیرید. برای این مقادیر با استفاده از مدل 5 دسته‌بندی به اجرا درآمده و نتایج در قالب شکل (3) نشان داده شده است. توجه شود برای رسم داده‌های فازی معادل قطعی شده آنها از روش درجه عضویت حداکثر-میانگین یعنی $(a + b)/2$ استفاده شده است.

$$minimize\ \theta^t$$

Subject to:

$$\sum_{i=1}^{n} \lambda_i x_{ji}^L \le \theta^t x_{jt}^L, \quad j = 1, \dots, m$$

$$\sum_{i=1}^{n} \lambda_i x_{ji}^R \le \theta^t x_{jt}^R, \quad j = 1, \dots, m$$

$$\sum_{i=1}^{n} \lambda_i x_{ji}^L - \sum_{i=1}^{n} \lambda_i \propto_{ji}^L \le \theta^t x_{jt}^L - \theta^t \propto_{jt}^L, j = 1, \dots, m$$

$$\sum_{i=1}^{n} \lambda_i x_{ji}^R + \sum_{i=1}^{n} \lambda_i \propto_{ji}^R \le \theta^t x_{jt}^R - \theta^t x_{jt}^R, \quad j = 1, \dots, m$$

$$\sum_{i=1}^{n} \lambda_i y_{ri}^L \ge y_{rt}^L, \quad r = 1, \dots, s$$

$$\sum_{i=1}^{n} \lambda_i y_{ri}^R \ge y_{rt}^R, \quad r = 1, \dots, s,$$

$$\sum_{i=1}^{n} \lambda_i y_{ri}^L - \sum_{i=1}^{n} \lambda_i \beta_{ri}^L \le y_{rt}^L - \beta_{rt}^L, \quad r = 1, \dots, s$$

$$\sum_{i=1}^{n} \lambda_i y_{ri}^R + \sum_{i=1}^{n} \lambda_i \beta_{ri}^R \ge y_{rt}^R - \beta_{rt}^R, \quad r = 1, \dots, s$$

$$\sum_{i=1}^{n} \lambda_i = 1,$$

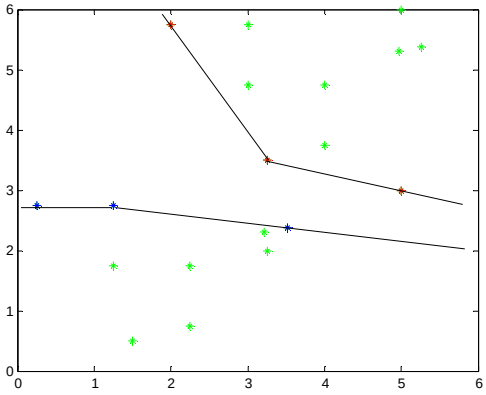
$$\lambda_i \ge 0, \quad i = 1, \dots, n \qquad (5)$$

جدول (5): مقادیر متغیر وابسته θ در حل سری مسائل LP فازی مربوط به دسته 1

شماره داده (دسته **1**)	θ
1	1.00
2	0.94
3	0.87
4	1.00
5	0.82
6	1.00
7	0.69
8	0.95
9	0.74
10	0.72

جدول (6): مقادیر متغیر وابسته θ در حل سری مسائل LP فازی مربوط به دسته 2

شماره داده (دسته **2**)	θ
1	1.67
2	1.31
3	1.00
4	1.00
5	1.25
6	1.06
7	1.06
8	1.34
9	1.03
10	1.00



شکل (3): محدوده ی دسته‌های 1 و 2 بر اساس معادل قطعی شده داده‌های فازی و نقاط مرزی

4-مدل جدید دسته‌بندی داده‌های جریان فازی با استفاده از تحلیل پوششی داده‌ها

مدل جدید ارائه شده در این تحقیق به منظور دسته‌بندی داده‌های فازی با استفاده از تحلیل پوششی داده در شکل (4) نشان داده شده است. همانطور که در شکل نشاد داده شده است، ابتدا بر مبنای داده‌های آموزشی که تعدادی یا همه مشخصه های آنها بصورت فازی ارزش گذاری شده اند و همچنین برچسب هر داده نیز در آن مشخص است، مسائل DEA به منظور شناسائی نقاط مرزی حل می شوند. از نقاط مرزی بدست آمده در هر کدام از مسائل DEA برای تعیین معیار تعلق داده‌ها به آن دسته استفاده خواهد شد. قبل از شروع دسته‌بندی جریان داده، متغیر D را تعریف و برابر صفر قرار می دهیم. از این متغیر در روند دسته‌بندی داده‌های جریانی برای نگهداری میزان فاصله داده‌هایی که مدل دسته‌های شناسائی شده را تائید نمی کنند استفاده خواهد شد. همچنین مجموعه $\mathbf{S_{new}}$ را تعریف و آن را برای نگهداری داده‌هایی که مدل دسته‌های شناسائی شده را تائید نمی کنند استفاده خواهد شد. این مجموعه در شروع تهی است.

در دسته‌بندی داده‌هایی که بصورت جریانی وارد می‌شود، چنانچه یک داده جدید، y، با توجه به محدوده های شناسائی شده جاری به یکی از

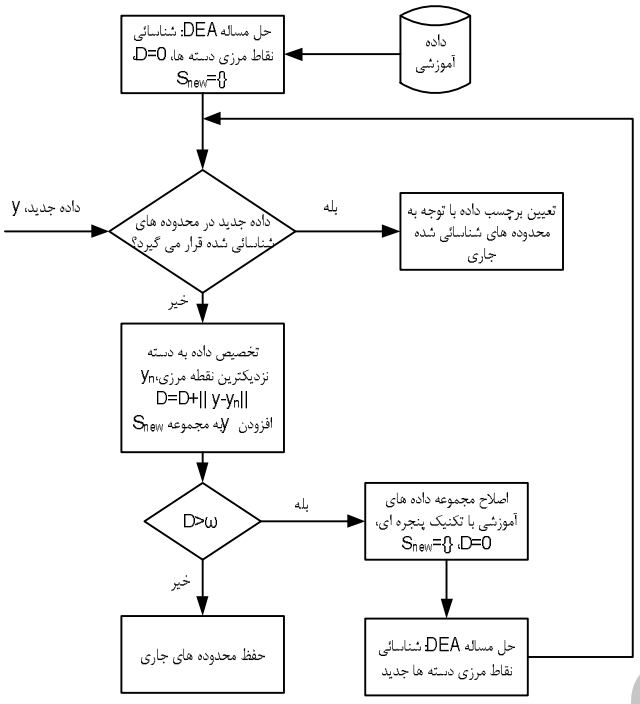
دسته‌ها تعلق گیرد، برچسب داده جدید را متناسب با دسته ای که به آن تعلق دارد، تعیین می‌کنیم. در غیر این صورت برچسب داده را برابر با برچسب نزدیکترین نقطه مرزی از نقاط مرزی جاری، y_n، قرار می دهیم.
برای محاسبه فاصله دو نقطه که معادل دو عدد فازی ذوزنقه ای می باشد از رابطه زیر استفاده می کنیم [3].

$$d(\tilde{x}_1, \tilde{x}_2) = \sqrt{\frac{1}{6}\left[\left((a_1 - c_1) - (a_2 - c_2)\right)^2 + 2(a_1 - a_2)^2 + 2(b_1 - b_2)^2 + \left((b_1 + d_1) - (b_2 + d_2)\right)^2\right]} \quad (6)$$

در همین حال، فاصله بین y و y_n را به D افزوده و داده y را به مجموعه S_{new} اضافه می کنیم. چنانچه مقدار D از حد آستانه از پیش تعیین شده ω کمتر باشد، محدوده دسته‌ها برای داده بعدی بر اساس نقاط مرزی جاری تعیین خواهد شد. در غیر این صورت می‌توان استنباط کرد محدوده های فعلی بصورت معناداری توانایی پوشش تمامی داده‌های را ندارند. یعنی سیستم دچار تغییراتی در زمان گردیده است و این تغییر در سیستم در قالب تغییر در الگو رفتار داده‌ها تجلی پیدا نموده است. لذا نیاز است محدوده های قبلی اصلاح شوند.
برای اصلاح محدوده ها، با توجه که مبنا نقاط مرزی در حل مساله های DEA است، مسائل DEA جدیدی برمبنای داده‌های جدید باید حل شوند. این مجموعه داده جدید از افزودن اعضای S_{new} به مجموعه داده آموزشی جاری و حذف همان تعداد از اعضای قدیمی تر مجموعه آموزشی جاری بدست می آید. با این اقدام که تکنیک پنجره ای[4] نام دارد، تعداد اعضای مجموعه داده آموزشی ثابت باقی مانده و فرصت برای ایجاد محدوده های جدید که بتواند دسته‌بندی صحیح را بر اساس وضعیت سیستم انجام دهد، ایجاد می‌گردد. پس از این اقدامات و اصلاح مجموعه داده‌های آموزشی نقاط جدید مرزی بدست آمده و محدوده های دسته‌های جدید قابل شناسائی می‌گردد. در این مرحله همچنین مقدار D را برابر صفر و S_{new} را خالی می کنیم. در این می‌تواند داده جدید بعدی را با محدوده های به هنگام شده دسته‌بندی نمود. این اقدامات تا زمانی که جریان داده وجود داشته باشد ادامه پیدا خواهد نمود.

5-آزمون مدل جدید دسته‌بندی جریان داده فازی با استفاده از تحلیل پوششی داده

برای تولید داده‌های آموزشی اولیه و همچنین داده‌هایی که بصورت جریانی به مدل پیشنهادی تغذیه می‌شود از [11] اقتباس می‌شود. البته در این مقاله تنها دسته‌بندی داده‌های فازی در حالت ایستا مد نظر بوده است. بدین منظور برای تولید 40 داده‌های آموزشی اولیه (دو بعدی) بصورت جداول (7) و (8) عمل می‌شود. برای ایجاد داده‌هایی که بصورت جریانی وارد می شوند پس از ایجاد تعداد 40 داده با استفاده از روش اشاره شده در جداول 9 و 10، به تعداد 180 داده فازی در بصورت تصادفی در فاصله بین این دو گروه از داده‌ها ایجاد می کنیم. در واقع در مجموع 40 داده آموزشی اولیه (جدول (12)) و 220 داده که بصورت جریانی وارد می شوند، تولید می‌شود.



شکل (4): مدل جدید دسته‌بندی جریان داده فازی با استفاده از DEA

جدول(7): شیوه ایجاد داده‌های فازی دسته 1 به منظور استفاده در مرحله آموزش اولیه

$a \sim N(7,1.5)$ $b' \sim U(0,1)$ $b = a + b'$ $c, d \sim U(0.2,0.7)$	بعد اول (مشخصه 1)
$a \sim N(9,1.5)$ $b' \sim U(0,1)$ $b = a + b'$ $c, d \sim U(0.2,0.7)$	بعد دوم (مشخصه 2)

جدول(8): شیوه ایجاد داده‌های فازی دسته 2 به منظور استفاده در مرحله آموزش اولیه

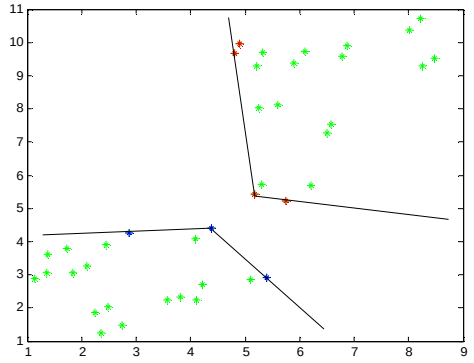
a~N(3,1.5) b′~U(0,1) b = a + b′ c, d~U(0.2,0.7)	بعد اول (مشخصه 1)
a~N(3,1.5) b′~U(0,1) b = a + b′ c, d~U(0.2,0.7)	بعد دوم (مشخصه 2)

جدول (9): شیوه ایجاد داده‌های فازی دسته 1 به منظور استفاده در مرحله

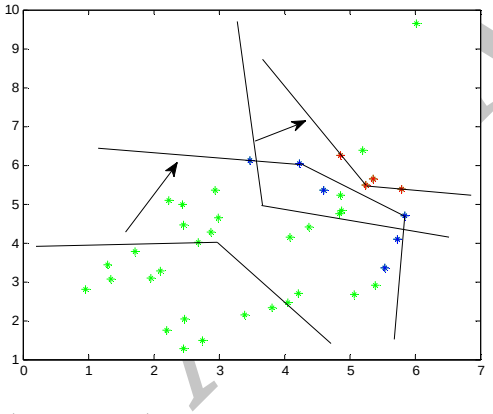
[4]- Windowing Technique

Recall	100	76/67	100	92/22
F_1	96/26	86/79	89/95	95/95

حجم داده‌های خارج از محدوده خوشه ها در شکل (7) نمایش داده شده است. لازم به ذکر است برای حل مدل های برنامه ریزی خطی که در حل مسائل DEA با آن مواجه می شویم از نرم افزار YALMIP که در MATLAB اجرا می‌گردد، استفاده شده است.



شکل (5): محدوده های دسته‌های 1 و 2 بدست آمده از داده‌های آموزشی اولیه



شکل (6): محدوده های دسته‌های 1 و 2 بدست آمده از داده‌های آموزشی اولیه و پس از ورود 220 داده جدید

آخر تست

بعد اول (مشخصه 1)	a~N(8,1.5) b′~U(0,1) b = a + b′ c, d~U(0.2,0.7)
بعد دوم (مشخصه 2)	a~N(9,1.5) b′~U(0,1) b = a + b′ c, d~U(0.2,0.7)

جدول(10): شیوه ایجاد داده‌های فازی دسته 2 به منظور استفاده در مرحله آخر تست

بعد اول (مشخصه 1)	a~N(2,1.5) b′~U(0,1) b = a + b′ c, d~U(0.2,0.7)
بعد دوم (مشخصه 2)	a~N(4,1.5) b′~U(0,1) b = a + b′ c, d~U(0.2,0.7)

پس از اجرا روش جدید دسته‌بندی جریان داده فازی با استفاده از DEA در محیط نرم افزار MATLAB، از شاخص‌های صحت[5] و F_1 که بصورت زیر تعریف می شوند، برای ارزیابی روش پیشنهادی استفاده می‌گردد:

$$\text{صحت} = \frac{(\text{تعداد داده ها یی که دسته آنها درست پیش بینی شده})}{\text{تعداد کل داده ها}} \qquad (7)$$

$$F_1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (8)$$

دقت(Precision) = مدل چه نسبتی از داده‌های دسته مورد نظر را شامل می‌شود

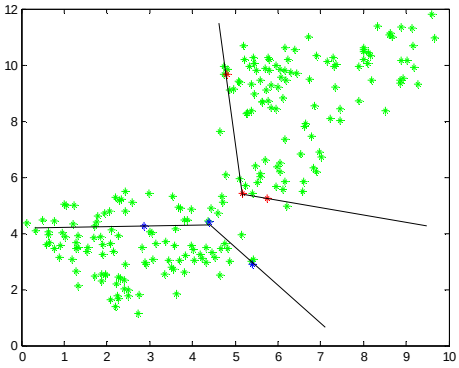
بازیابی(Recall) = نسبتی از داده‌هایی که به درستی به دسته ای خاص تعلق گرفته

نتایج حاصل از بکار بردن روش پیشنهادی در جدول (11) ارائه شده است. در این جدول روش پیشنهادی با حالتی که بهنگام سازی در آن اجرا نمی‌شود، مقایسه گردیده است.

جدول (11): نتایج مقایسه روش پیشنهادی

	روش پیشنهادی(با بهنگام سازی)		روش پایه(بدون بهنگام سازی)	
	دسته اول	دسته دوم	دسته اول	دسته دوم
Precision	92/87	100	81/08	100

[5] - Accuracy



شکل (7): داده‌های جدید قرار گرفته در خارج از محدوده اولیه دسته‌ها

جدول(12): داده‌های اولیه

ردیف	مشخصه 1				مشخصه 2				دسته
	a	b	c	d	a	b	C	d	
1	7.99998	8.504175	0.441703	0.388807	8.892574	9.697303	0.204586	0.522357	1
2	4.91105	5.430771	0.607884	0.304314	5.378114	5.490466	0.307851	0.654644	1
3	5.04915	5.437951	0.223071	0.352402	7.958476	8.103155	0.303739	0.62532	1
4	6.09246	6.906306	0.541047	0.641219	6.912916	7.626026	0.240212	0.60935	1
5	4.76715	4.848952	0.447413	0.415165	9.494472	9.84343	0.281274	0.469875	1
6	7.83781	8.183272	0.65683	0.280086	9.897817	10.82114	0.418562	0.470063	1
7	6.58396	6.971838	0.683608	0.233362	9.220763	9.919116	0.201027	0.618302	1
8	5.05947	5.3762	0.263362	0.2579	8.847842	9.737265	0.209876	0.24921	1
9	5.66734	5.831705	0.246974	0.466066	5.047529	5.421684	0.683739	0.509793	1
10	5.52022	6.282624	0.408748	0.617765	9.04208	9.712118	0.428085	0.619018	1
11	8.41497	8.666953	0.385917	0.4411	9.151455	10.02969	0.467666	0.468209	1
12	5.15700	5.727674	0.37389	0.449352	5.693027	5.876463	0.493079	0.456103	1
13	5.50768	5.690969	0.361863	0.457605	8.379598	8.249429	0.518492	0.443997	1
14	6.49247	6.962891	0.349836	0.465858	7.376944	7.89972	0.543905	0.431891	1
15	5.09917	4.977806	0.33781	0.474111	9.808823	10.20422	0.569317	0.419785	1
16	7.93184	8.299742	0.325783	0.482364	10.32963	11.14334	0.59473	0.407678	1
17	7.07128	7.314199	0.313756	0.490617	9.64187	10.28555	0.620143	0.395572	1
18	5.51827	5.385751	0.301729	0.49887	8.956465	9.978333	0.645556	0.383466	1
19	6.12818	6.051681	0.289702	0.507123	5.289551	5.52915	0.670969	0.37136	1
20	5.64508	6.324646	0.277676	0.515375	9.289791	10.19117	0.696382	0.359254	1
21	1.68553	2.508633	0.471056	0.246196	2.961103	3.586489	0.456646	0.487854	2
22	2.60178	2.89706	0.501119	0.291419	1.334058	1.621302	0.646342	0.246296	2
23	2.50863	3.241803	0.474446	0.27808	4.126251	4.418736	0.610949	0.69701	2
24	1.26262	2.187699	0.301078	0.204224	3.750251	3.818067	0.686543	0.661394	2
25	3.87008	4.559191	0.400484	0.389791	2.224109	3.18284	0.481582	0.675263	2
26	3.35963	4.26282	0.671675	0.43699	2.161186	2.47941	0.332949	0.230542	2
27	2.47367	2.491433	0.609977	0.228694	1.869944	2.202712	0.346211	0.401242	2
28	4.33814	4.428276	0.548327	0.273575	4.388719	4.422831	0.343734	0.309811	2
29	5.36744	5.425027	0.231679	0.561415	2.62722	3.193147	0.635046	0.327493	2

30	1.33773	1.371447	0.547859	0.310588	2.775248	3.367059	0.461346	0.272433	2
31	1.39333	2.137378	0.43719	0.41345	2.71737	3.281544	0.513742	0.235558	2
32	2.12033	2.864583	0.418769	0.43079	1.018398	1.485687	0.536675	0.198684	2
33	2.18379	3.070627	0.400347	0.44813	3.885626	4.325432	0.559607	0.161809	2
34	1.08676	2.030524	0.381925	0.46547	3.48406	3.697166	0.58254	0.124935	2
35	3.42882	4.358029	0.363503	0.48281	2.105078	2.760174	0.605472	0.08806	2
36	3.21022	4.205422	0.345081	0.50015	2.06128	2.258857	0.628405	0.051186	2
37	2.31261	2.190214	0.326659	0.51749	1.812232	2.077177	0.651337	0.014311	2
38	4.17818	4.342631	0.308238	0.53483	4.348913	4.205796	0.67427	-0.02256	2
39	5.21753	4.943112	0.289816	0.55217	2.392648	3.177612	0.697203	-0.05944	2
40	1.28225	1.22496	0.271394	0.56951	2.280905	3.300878	0.720135	-0.09631	2

6- نتیجه گیری

در این تحقیق مدلسازی مساله دسته‌بندی داده‌ها در قالب یک مساله تحلیل پوششی داده صورت گرفت. سپس مدلسازی به سطح یک مدل غیرقطعی فازی توسعه پیدا نمود تا بتوان داده‌های فازی ذوزنقه‌ای را مورد مطالعه قرار داد. در گام بعدی مدل برنامه‌ریزی ریاضی فازی به یک مدل برنامه ریزی قطعی معادل تبدیل شد بگونه ای که با حل آن مقادیر تابع هدف (کارایی‌DMUها) بصورت قطعی حاصل شود. با ارائه یک چارچوب جدید، مدل برنامه‌ریزی ایجاد شده جهت دسته‌بندی داده‌های جریانی فازی به کارگرفته شد. این چارچوب پیشنهادی این امکان را فراهم می آورد تا با ورود داده‌ها در طول زمان، به‌هنگام‌سازی محدوده‌های دسته‌ها اتفاق بیافتد.

به بکارگیری روش پیشنهادی جدید بر روی دادهای شبیه‌سازی شده (مستخرج از مقالات موجود در ادبیات تحقیق)، صحت پیش‌بینی در دسته‌بندی به بیش از 96درصد رسید در صورتی‌که عدم استفاده از ویژگی به‌هنگام‌سازی دقت در حدود 86درصد بود. این درحالی است که میانگین شاخص F_1 نیز 10درصد بهبود داده شد. از طرفی روش پیشنهادی فقط نیاز به تعیین یک پارامتر ورودی توسط کاربر، یعنی پارامتر ε دارد و از این منظر کار را برای استفاده از آن ساده می کند.

7- منابع و مأخذ

[1] غضنفری، مهدی، علیزاده، سمیه، تیمورپور، بابک، (1387)، **داده کاوی و کشف دانش**، دانشگاه علم و صنعت ایران.

[2] Abonyi, J., Feil, B., (2007), **Cluster analysis for data mining and system identification**, Springer Science & Business Media.

[3] Chen, T.Y., Ku, T.C., Tsui, C. W., (2008), **Determining attribute importance based on triangular and trapezoidal fuzzy numbers in (z) fuzzy measures**, In The 19th International Conference on Multiple Criteria Decision Making (pp. 75-76).

[4] Jain, A.K., (2010), **Data clustering: 50 years beyond K-means**, Pattern recognition letters, 31(8), 651-666.

[5] Lee, S., Kim, G., Kim, S., (2011), **Self-adaptive and dynamic clustering for online anomaly detection**, Expert Systems with Applications, 38(12), 14891-14898.

[6] León, T., Liern, V., Ruiz, J. L., Sirvent, I., (2003), **A fuzzy mathematical programming approach to the assessment of efficiency with DEA models**, Fuzzy sets and systems, 139(2), 407-419.

[7] Mena-Torres, D., Aguilar-Ruiz, J.S., (2014), **A similarity-based approach for data stream classification**, Expert Systems with Applications, 41(9), 4224-4234.

[8] Pendharkar, P.C., Troutt, M.D., (2014), **Interactive classification using data envelopment analysis**, Annals of Operations Research, 214(1), 125-141

[9] Taneja, S., Suri, B., Narwal, H., Jain, A., Kathuria, A., Gupta, S., (2016), **A new approach for data classification using Fuzzy logic**, In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence) ,pp. 22-27, IEEE.

[10] Yan, H., Wei, Q., (2011),**Data envelopment analysis classification machine**, Information Sciences, 181(22), 5029-5041.

[11] Yazdi, H. S., Vhedian, A., (2009), **Fuzzy Bayesian classification of LR Fuzzy numbers**, IACSIT International Journal of Computer Theory and Engineering, 1(5).

[12] Yin, S., Kaynak, O., (2015), **Big Data for Modern Industry: Challenges and Trends [Point of View]**, Proceedings of the IEEE, 103(2), 143-146.