به نام خدا

موضوع

ارایه روشی جدید برای انتخاب خوشه بندی ترکیبی با استفاده از معیارهای پراکندگی و استقلال

نویسندگان

محمد یوسف نژاد[۱]، حسین علیزاده[۲]، بهروز مینایی بیدگلی[۲]

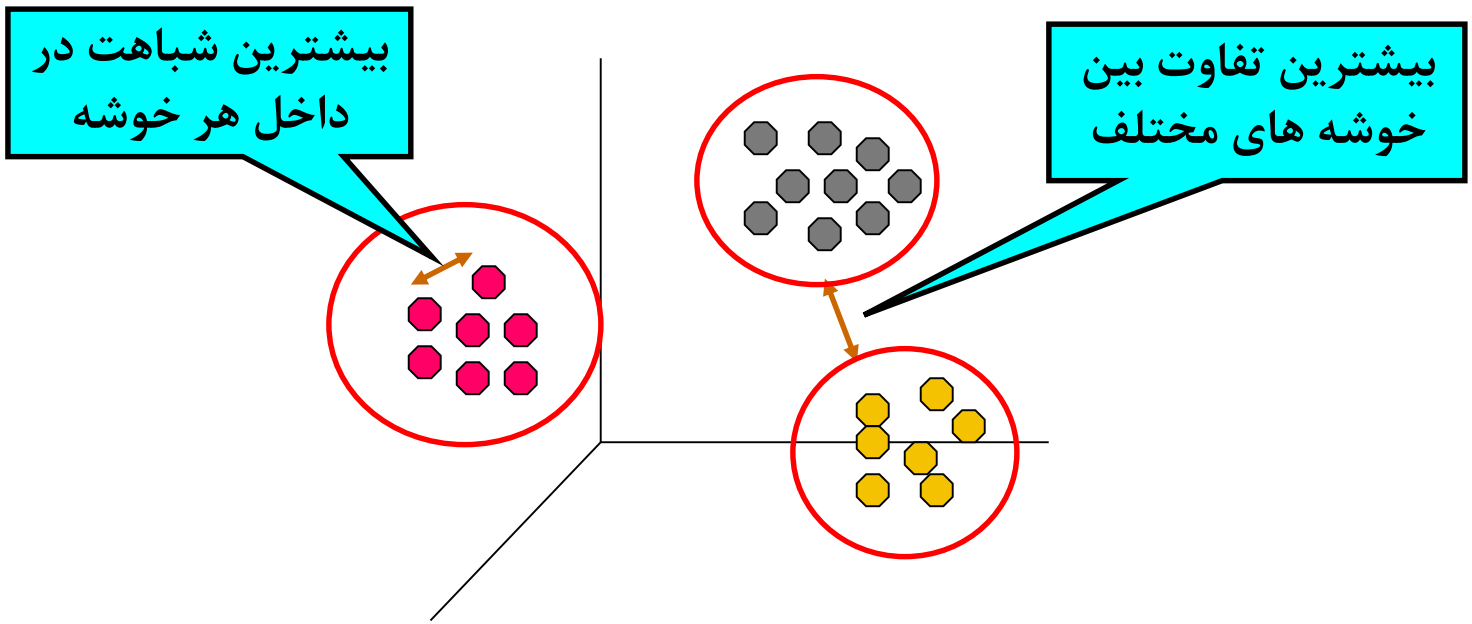
[۱] دانشکده فناوری اطلاعات، دانشگاه علوم و فنون مازندران

[۲] دانشکده کامپیوتر، دانشگاه علم و صنعت ایران

فهرست

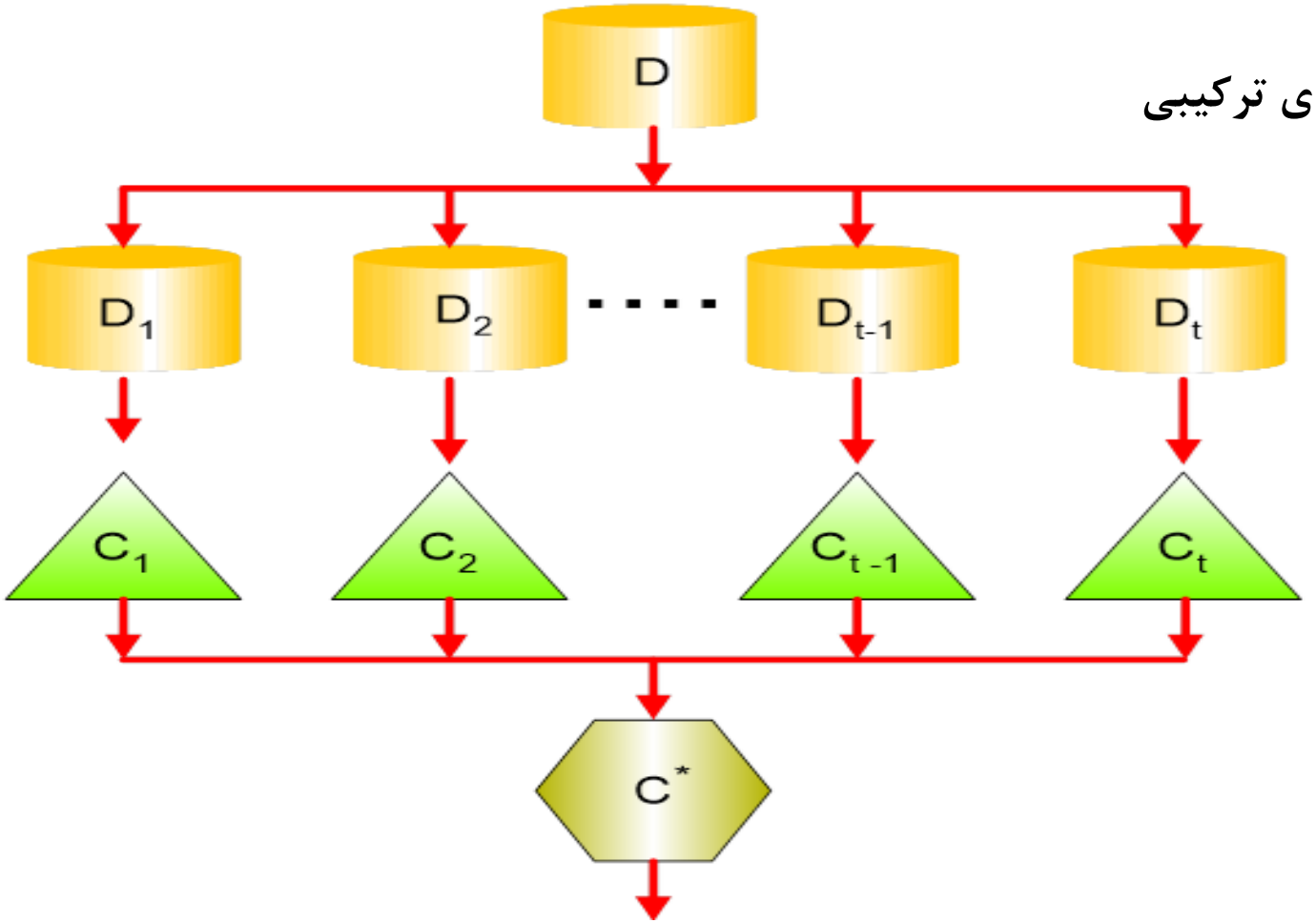
خوشه بندی

- خوشه بندی یک روش **بدون ناظر** است.
- در خوشه بندی نمونه‌ها بر اساس **معیار شباهت/تفاوت** جدا می شوند.
- نتایج الگوریتم‌های پایه خوشه بندی نسبت به معیار شباهت/تفاوت **بهینه** خواهند شد.
  - ✓ خوشه‌بندی پایه روی جنبه‌های خاصی از داده‌ها تاکید می‌کنند.



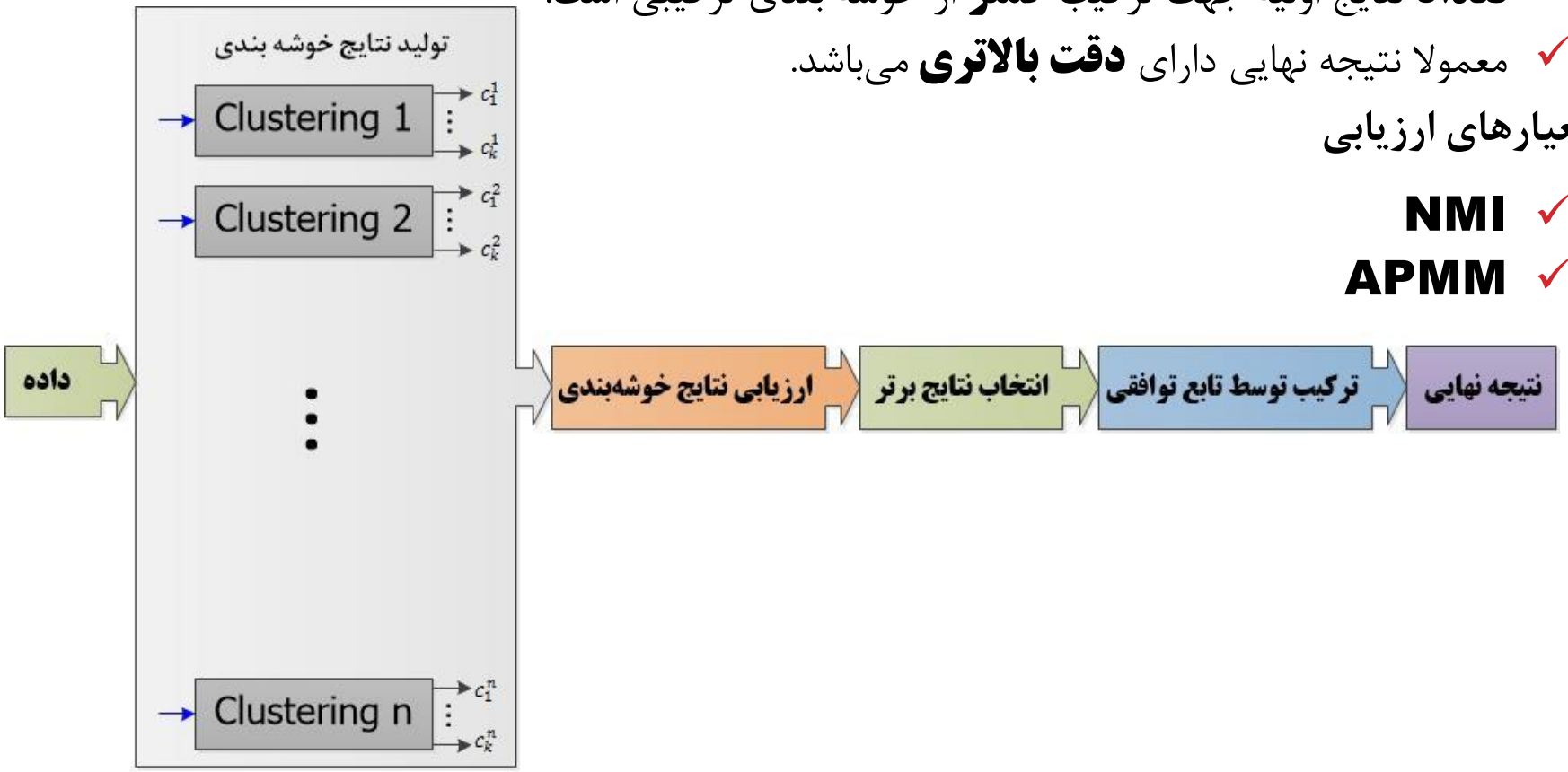
خوشه بندی ترکیبی

- **خوشه‌بندی ترکیبی نتایج به دست آمده از خوشه‌بندی‌های اولیه را جهت تولید خوشه نهایی ترکیب می کند.**
- **ویژگی‌های خوشه‌بندی ترکیبی**
  - ✓ استحکام
  - ✓ نو بودن
  - ✓ پایداری
  - ✓ انعطاف پذیری



خوشه بندی ترکیبی مبتنی بر انتخاب

- یافت **افرازها/خوشه‌های** موثرتر بر اساس **معیار ارزیابی** و ترکیب آن‌ها
- ویژگی‌های اصلی
  - ✓ **تعداد** نتایج اولیه جهت ترکیب **کمتر** از خوشه بندی ترکیبی است.
  - ✓ معمولا نتیجه نهایی دارای **دقت بالاتری** می‌باشد.
- معیارهای ارزیابی
  - ✓ **NMI**
  - ✓ **APMM**

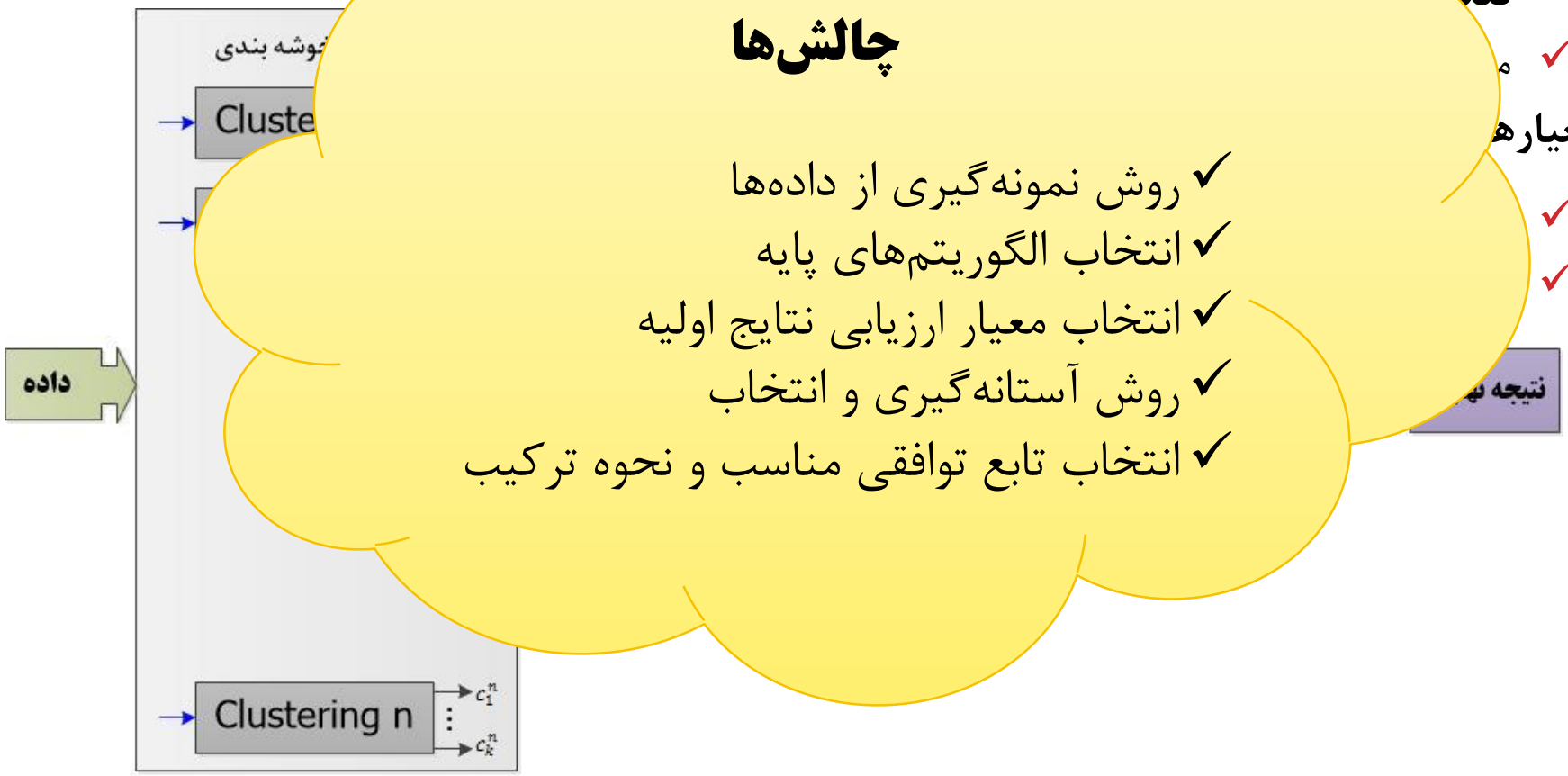


خوشه بندی ترکیبی مبتنی بر انتخاب

- یافت **افرازها/خوشه‌های** موثرتر بر اساس **معیار ارزیابی** و ترکیب آن‌ها
- ویژگی‌های اصل
  - ✓ **تعدا**
  - ✓ م
- معیارها
  - ✓
  - ✓

چالش‌ها

- ✓ روش نمونه‌گیری از داده‌ها
- ✓ انتخاب الگوریتم‌های پایه
- ✓ انتخاب معیار ارزیابی نتایج اولیه
- ✓ روش آستانه‌گیری و انتخاب
- ✓ انتخاب تابع توافقی مناسب و نحوه ترکیب



روش پیشنهادی مقاله

- در این روش از **کل داده** برای تمام الگوریتم‌های پایه استفاده می‌شود.
- الگوریتم‌های پایه به صورت اختیاری می‌تواند **هر نوع الگوریتم ساده یا ترکیبی** باشد.
- معیار ارزیابی نتایج اولیه بر اساس **پراکندگی و استقلال** می‌باشد.
- آستانه‌گیری بر اساس **زمان اجرای الگوریتم** برای دو معیار انتخاب می‌شود.
- تابع توافقی این مقاله روش انباشت مدارک (EAC) می باشد.

$$C(i, j) = \frac{n_{i,j}}{m_{i,j}}$$

معیار پراکندگی

- پراکندگی یک **معیار کلاسیک** جهت ارزیابی خوشه و خوشه بندی می باشد.
- معروف‌ترین معیار جهت ارزیابی پراکندگی **NMI** است.
- معیار **NMI مشکل تقارن** دارد.
- معیار **APMM** معیاری بر اساس **NMI** برای ارزیابی **خوشه** است.
- در این مقاله معیاری جدید بر اساس معیار **APMM** با عنوان **A3** معرفی شده است.
- **A3** می‌تواند **افراز را ارزیابی** کند.

$$AAPMM(C_i) = \frac{1}{M}\sum_{j=1}^{M} APMM\left(C_i, P_j^{b*}\right)$$

$$A3(P) = \frac{1}{n}\sum_{i=1}^{k} n_i \times AAPMM(C_i)$$

- شرط پراکندگی

$$Diversity(P) \geq dT$$

معیار استقلال

- در **داده‌های پیچیده** همواره پراکندگی خوشه معیار افرازبندی خوبی نیست.
- استقلال الگوریتم احتمال درستی نتایج اولیه را بر اساس نوع الگوریتم بررسی می‌کند.
- در روش پیشنهادی دو الگوریتم غیر هم نام **کاملا مستقل** در نظر گرفته می‌شوند.
- استقلال الگوریتم‌های پایه هم نام بر اساس **مقادیر پایه** آن‌ها محاسبه می‌شود.

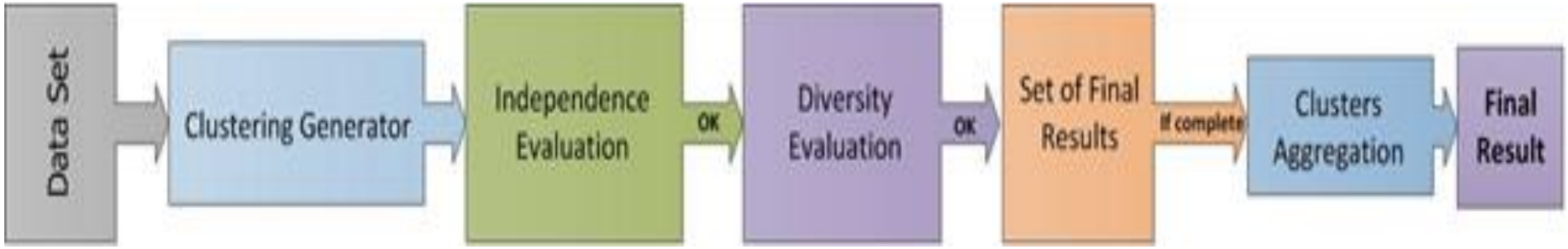
روش محاسبه استقلال افزازها

$$Independence(P) = \frac{1}{M}\sum_{i=1}^{M} \mathrm{BIndependence}(P, P_i)$$

شرط استقلال

$$Independence(P) \geq iT$$

روند اجرای الگوریتم پیشنهادی



ارزیابی – مجموعه داده‌های (استاندارد)

No.	Name	Feature	Class	Sample
1	Half Ring	2	2	400
2	Iris	4	3	150
3	Balance Scale	4	3	625
4	Breast Cancer	9	2	683
5	Bupa	6	2	345
6	Galaxy	4	7	323
7	Glass	9	6	214
8	Ionosphere	34	2	351
9	SA Heart	9	2	462
10	Wine	13	2	178
11	Yeast	8	10	1484
12	Pendigits	16	10	10992
13	Statlog	36	7	6435
14	Optdigits	62	10	5620

ارزیابی - نتایج

	روش های خوشه بندی پایه				روش های خوشه بندی ترکیبی					روش پیشنهادی مقاله		
	Kmeans	FCM	Subtract	Single Linkage	EAC	MAX	CSPA	HGPA	MCLA	*iT*	*dT*	نتیجه
Half Ring	75.75	78	86	75.75	77.17	78.48	74.5	50	74.5	0.2	0.06	**87.2**
Iris	65.3	82.66	55.3	68	**96**	72.89	85.34	48.66	89.34	0.2	0.06	**96**
Balance Scale	40.32	44	45.32	46.4	52	52.1	51.84	41.28	51.36	0.23	0.063	**54.88**
Breast Cancer	93.7	94.43	65	65.15	95.02	75.72	80.97	50.37	96.05	0.18	0.02	**96.92**
Bupa	54.49	50.1	**57.97**	57.68	55.18	56.17	56.23	50.32	55.36	0.21	0.04	57.42
Galaxy	30.03	34.98	29.72	25.07	31.95	32.78	29.41	31.27	28.48	0.2	0.05	**35.88**
Glass	42.05	47.19	36.44	36.44	45.93	44.17	38.78	41.12	51.4	0.19	0.06	**51.82**
Ionosphere	69.51	67.8	**71.5**	64.38	70.48	64.48	67.8	58.4	70.22	0.3	0.1	70.52
SA Heart	64.51	63.41	67.26	65.15	65.19	63.96	58.42	50.93	62.54	0.65	0.8	**68.7**
Yeast	31.19	29.98	31.2	31.73	31.74	32.4	14	15.23	17.56	0.5	0.5	**34.76**
Wine	65.73	**71.34**	67.23	37.64	70.56	69.17	67.41	62.36	70.22	0.2	0.05	**71.34**
Pendigits	46.97	36.77	10.4	10.46	10.47	57.02	58.32	11.14	58.62	0.02	0.12	**58.68**
Optdigit	52.52	38.33	47.72	10.28	20	76.11	75.21	64.77	77.15	0.01	0.1	**77.16**
Statlog	50.93	49.91	23.8	23.8	23.9	54.23	54.23	52.94	55.71	0.01	0.1	**55.77**

- **به طور متوسط روش پیشنهادی بیشتر از ۱۰ درصد نتایج را بهبود داده است.**

نتیجه گیری

- **مزایا**
  - ✓ برای اولین بار مسئله تاثیر **استقلال الگوریتم‌ها** در نتیجه نهایی خوشه بندی مطرح شده است.
  - ✓ در این روش به جای توجه به بیشینه سازی پراکندگی در مسئله، **احتمال درستی تمام الگوها** بررسی می‌شود.
  - ✓ رفع **مشکل تقارن** در معیار پراکندگی با توسعه معیار APMM برای ارزیابی افرازبندی‌ها
  - ✓ امکان استفاده از هر نوع الگوریتم به عنوان الگوریتم خوشه بندی اولیه
- **معایب**
  - ✓ این روش نیاز به **آستانه‌گیری** برای هر دو معیار پراکندگی و استقلال دارد.
  - ✓ این روش الگوریتم‌های غیرهم نام را **کاملا مستقل** در نظر می‌گیرد.
- **کارهای آینده**
  - ✓ حل معایب فعلی الگوریتم پیشنهادی
  - ✓ توجه به سایر پارامترهای موثر در مسئله خوشه‌بندی ترکیبی مبتنی بر انتخاب
    - بهینه‌سازی روند ترکیب نتایج اولیه
    - توجه ویژه به داده و ویژگی‌های آن

با تشکر از توجه شما

پایان