به نام خدا

خوشه‌بندی ترکیبی مبتنی بر زیرمجموعه‌ای از نتایج اولیه

حسین علیزاده

استاد راهنما: دکتر مینایی بیدگلی

اسفند ۱۳۸۷

رئوس مطالب

- مقدمه‌ای بر خوشه‌بندی ترکیبی
- روش پیشنهادی
  - ارزیابی خوشه
  - انتخاب خوشه
  - ساخت ماتریس همبستگی
- نتایج آزمایشات
- جمع‌بندی و کارهای آینده

خوشه‌بندی داده‌ها

- یک روش برای گروه‌بندی کردن نمونه‌ها در خوشه‌های شبیه به هم می‌باشد. به طوری که نمونه‌های هر خوشه **حداکثر تشابه را با یکدیگر** و **حداکثر فاصله را با نمونه‌های خوشه‌های دیگر** داشته باشند.

- یک شکل از یادگیری **بدون ناظر**، که برچسب رده نمونه ها مشخص نیست.

- یک روش جستجوی **الگوهای پنهان** در داده‌های بدون برچسب.

ضعف روش‌های پایه خوشه‌بندی

- روش‌های پایه خوشه‌بندی روی مجموعه داده‌های خاصی خوب عمل می کنند.
- مجموعه داده‌ها نیز ناشناخته هستند.

هیچ الگوریتم خوشه‌بندی وجود ندارد که **برای همه مجموعه داده‌ها** بهترین عملکرد را داشته باشد.

- دو سوال رایج برای یک مجموعه داده که:
  - بهترین خوشه‌بندی چیست؟
  - تعداد خوشه‌ها چیست؟

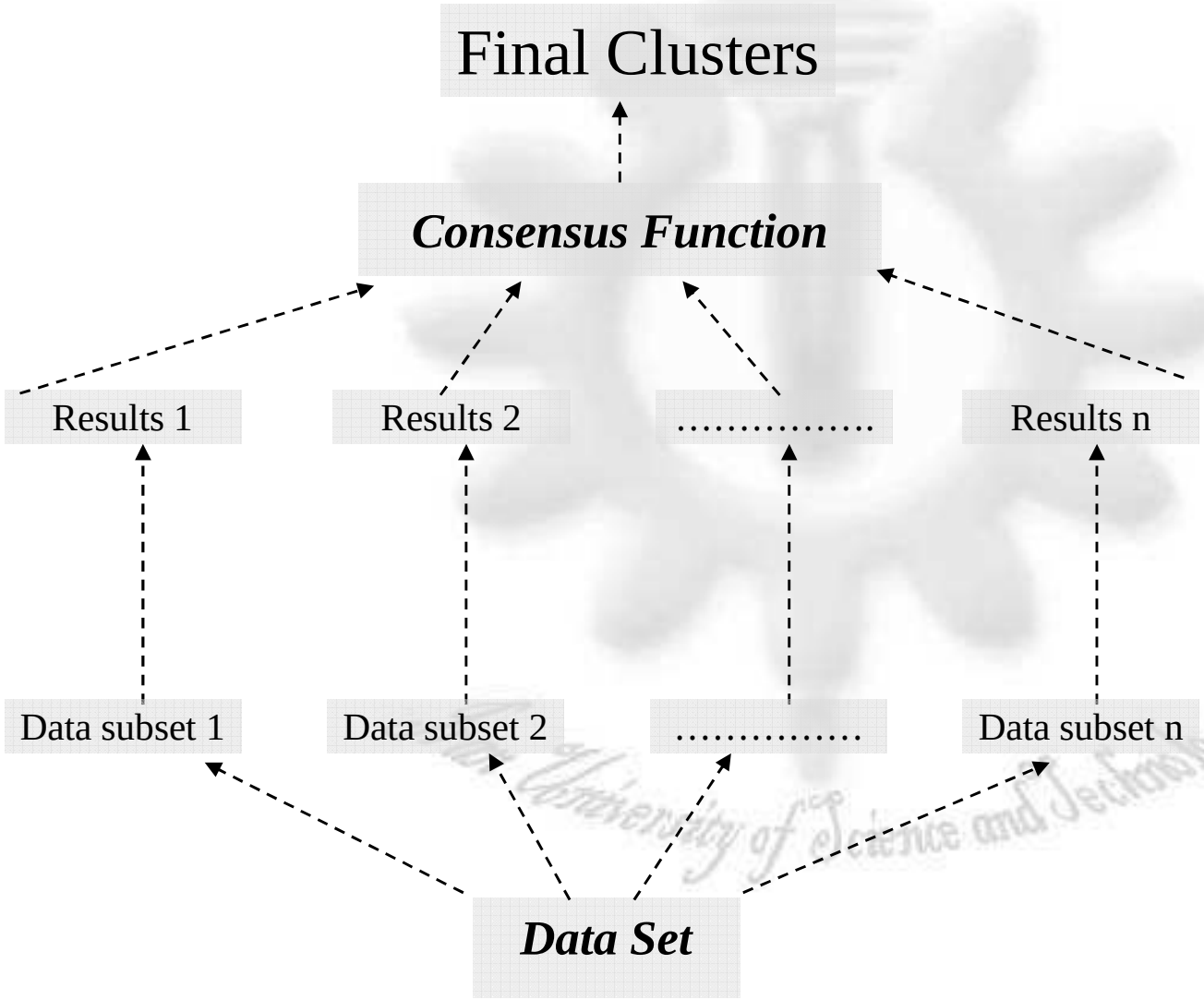
هیچ الگوریتم خوشه‌بندی وجود ندارد که بهترین راه‌حل برای هر دو سوال باشد.

خوشه‌بندی ترکیبی

- خوشه‌بندی ترکیبی، نتایج خوشه‌بندی‌های اولیه را برای به دست آوردن نتایج بهتر، با هم ترکیب می‌کند.

- نتایج خوشه‌بندی ترکیبی در مقابل خوشه‌بندی ساده
  - **استحکام** (Robustness)
  - **نو بودن** (Novelty)
  - **پایداری**(Stability)
  - **انعطاف‌پذیری** (Flexibility)

روش معمول خوشه‌بندی ترکیبی



استفاده از یک ترکیب کننده برای به دست آوردن نتیجه نهایی

به کار بردن الگوریتم‌های خوشه‌بندی با مشخصات مختلف

تولید زیرمجموعه‌های مختلف از داده‌ها


خوشه‌بندی ترکیبی
مکانیزم تولید
تابع توافقی (اجماع)
الگوریتم‌های مختلف
استفاده از یک الگوریتم
افرازبندی
سلسله‌مراتبی
مقداردهی اولیه مختلف
پارامترهای مختلف
تصویر داده‌ها به فضای ویژگی‌های جدید
زیرمجموعه‌های مختلف از ویژگی‌ها
زیرمجموعه‌های مختلف از داده‌ها
روش قطعی
بازنمونه‌برداری
روش‌های مبتنی بر ماتریس همبستگی
روش رای‌گیری
مدل آمیزشی
روش‌های مبتنی بر ابرگراف
Sing. Linkage
Avr. Linkage
Com. Linkage
CSPA
HGPA
MCLA

مفهوم پراکندگی

- پراکندگی در خوشه‌بندی ترکیبی به معنی وجود تفاوت و تمایز در نتایج اولیه می‌باشد.
- اکثر مطالعات اخیر
  - به کار گیری خوشه‌بندی‌های اولیه متنوع‌تر (پراکندگی بیشتر)

    جین، بوهمن، فرد، مینایی و تاپچی
- آیا پراکندگی به وجود آمده مفید می‌باشد یا نه؟
  - کارهای صورت گرفته توسط کانچوا نشان می‌دهد که ایجاد پراکندگی در خوشه‌بندی‌های اولیه معمولا موجب بهبود عملکرد خوشه‌بندی در اکثر مواقع می‌شود.
  - عظیمی (۱۳۸۶) نشان داده است که در بعضی مجموعه داده‌ها، پراکندگی بیشتر لزوما کمکی به افزایش دقت در نتایج نهایی نمی‌کند.

مفهوم کیفیت

- کانچوا و هاجیتودوروف
  - نشان می‌دهند که هر چه نتایج اولیه علاوه بر داشتن پراکندگی لازم، از کیفیت بالاتری برخوردار باشند، کیفیت خوشه‌های نهایی نیز بهتر خواهد بود .

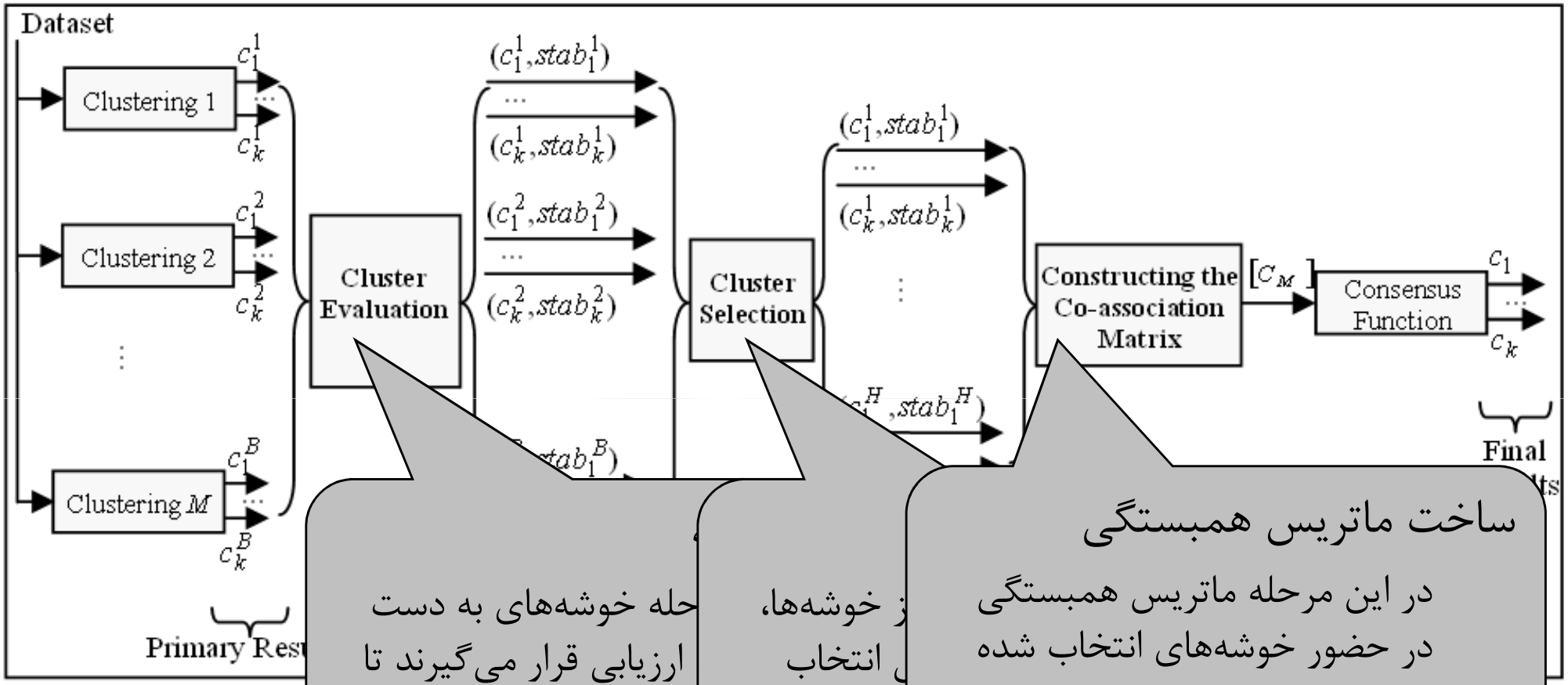
- فرن و لین (۲۰۰۸)
  - نشان داده‌اند که بهینه‌سازی همزمان دو عامل پراکندگی و کیفیت در نتایج اولیه می‌تواند کارایی خوشه‌بندی ترکیبی را به طور چشمگیری بهبود بخشد.

مفهوم پایداری

- یک خوشه پایدار، خوشه‌ای است که اگر آن روش خوشه‌بندی را چند بار دیگر هم، روی آن مجموعه داده (یا روی مجموعه‌های مختلف حاصل از نمونه‌برداری از آن مجموعه داده) اجرا کنیم، با احتمال زیاد این خوشه باز هم دیده خواهد شد.
- خوشه‌های پایدار به خوشه‌هایی اطلاق می‌شود که در خوشه‌بندی‌های مختلف روی زیرمجموعه‌های به دست آمده از نمونه‌برداری‌های مختلف بیشترین تکرار را داشته باشند.
- با تغییرات جزیی در مجموعه داده، آن خوشه باز هم تکرار شود.

- مقدمه‌ای بر خوشه‌بندی ترکیبی
- **روش پیشنهادی**
  - **ارزیابی خوشه**
  - **انتخاب خوشه**
  - **ساخت ماتریس همبستگی**
- نتایج آزمایشات
- جمع‌بندی و کارهای آینده

چهارچوب کلی روش پیشنهادی



گام اول: ارزیابی خوشه

- یکی از معیارهایی که می‌تواند به عنوان تابع برازندگی خوشه در نظر گرفته شود، معیار **پایداری** خوشه است.
- استفاده از پایداری برای ارزیابی خوشه اولین بار توسط **لانژ و همکاران، ۲۰۰۳** پیشنهاد گردید.
- معیارهای مورد استفاده برای ارزیابی خوشه
  - اطلاعات متقابل نرمال شده (NMI)
  - روش ماکزیمم (MAX)
  - روش AMM
  - روش اطلاعات متقابل نرمال اصلاح شده (ENMI)

اطلاعات متقابل نرمال شده (NMI)

- برای اندازه‌گیری میزان شباهت دو افراز مختلف از داده‌ها
- اطلاعات متقابل اولین بار توسط استرل گاش (۲۰۰۲)
- نسخه نرمال شده اولین بار توسط ف

تعداد نمونه‌های مشترک بین خوشه‌های C_i از افراز a و C_j از افراز b

$$NMI(P^a, P^b) = \frac{-2\sum_{i=1}^{k_a}\sum_{j=1}^{k_b} n_{ij}^{ab} \log\left(\frac{n_{ij}^{ab}.n}{n_i^a.n_j^b}\right)}{\sum_{i}^{k_a} n_i^a \log\left(\frac{n_i^a}{n}\right) + \sum_{j=1}^{k_b} n_j^b \log\left(\frac{n_j^b}{n}\right)}$$

تعداد کل نمونه‌های موجود در خوشه C_i از افراز a

تعداد کل نمونه‌های موجود در خوشه C_j از افراز b

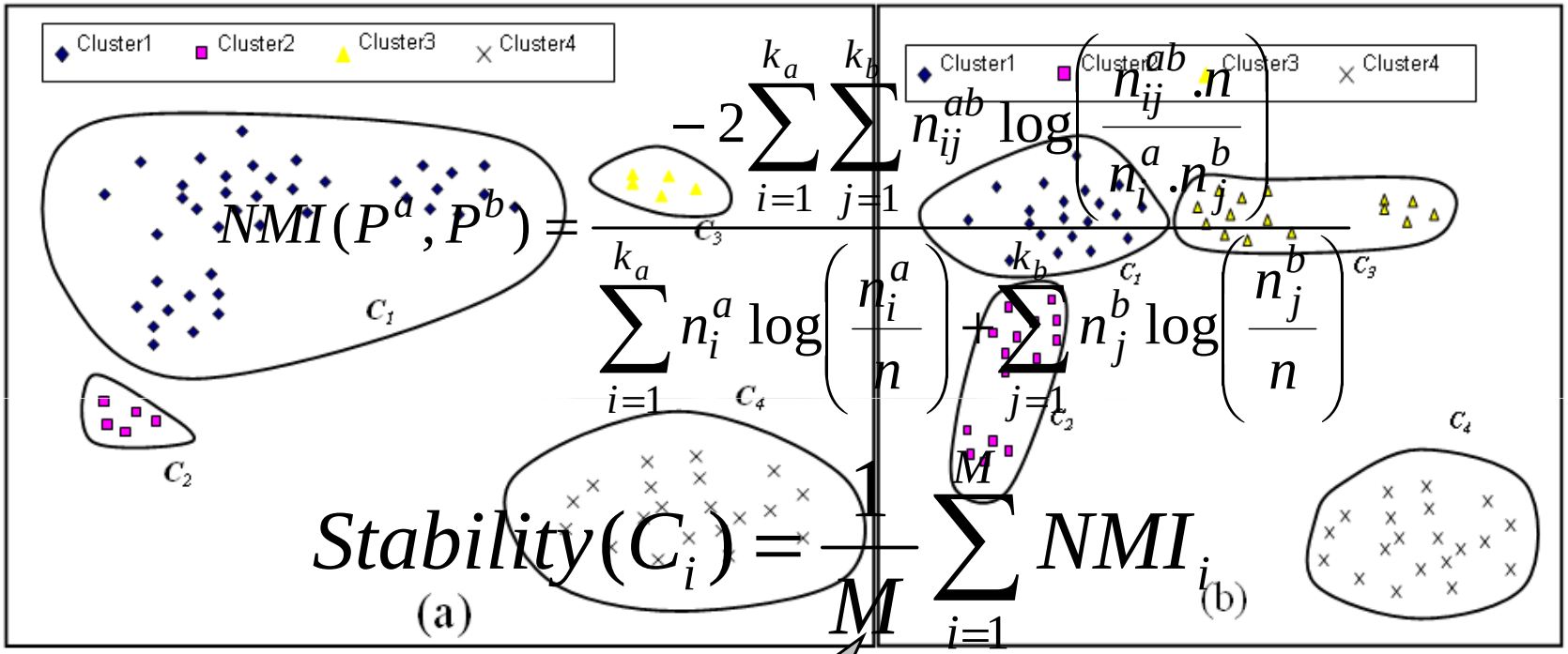
استفاده از رابطه NMI برای ارزیابی خوشه

- اولین بار توسط لانژ و همکاران، ۲۰۰۳
- یک مجموعه مرجع شامل تعدادی خوشه‌بندی اولیه در نظر می‌گیریم.

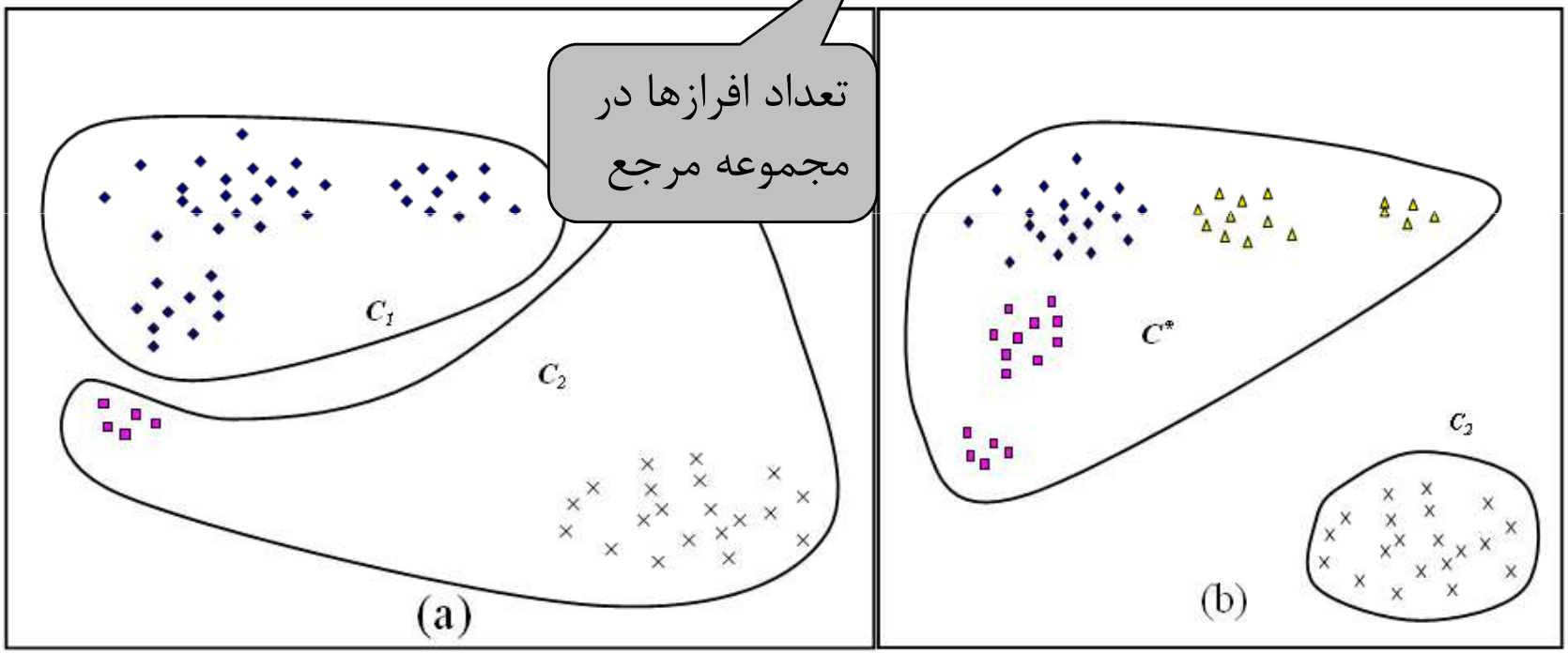
- روش کار برای ارزیابی یک خوشه نسبت به یک افراز:
  - تبدیل خوشه مورد نظر به افراز (بر اساس یک فرایند مشخص).
  - استفاده از رابطه NMI برای ارزیابی میزان شباهت بین بین این دو افراز
  - در نظر گرفتن نتیجه حاصل به عنوان میزان پایداری خوشه مورد نظر در افراز متناظر.

$$NMI(P^a, P^b) = \frac{-2\sum_{i=1}^{k_a}\sum_{j=1}^{k_b} n_{ij}^{ab} \log\left(\frac{n_{ij}^{ab}.n}{n_i^a.n_j^b}\right)}{\sum_{i=1}^{k_a} n_i^a \log\left(\frac{n_i^a}{n}\right) + \sum_{j=1}^{k_b} n_j^b \log\left(\frac{n_j^b}{n}\right)}$$

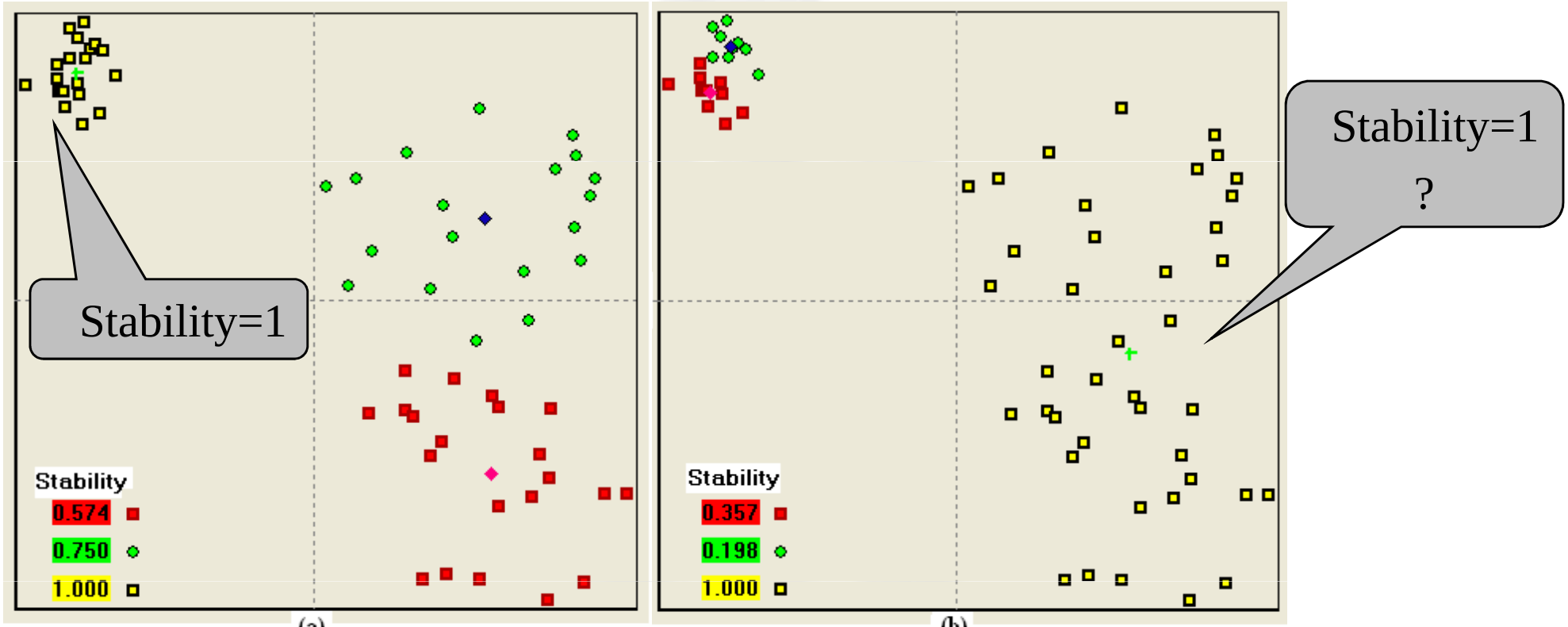
$$Stability(C_i) = \frac{1}{M}\sum_{i=1}^{M} NMI_i$$



تعداد افرازها در مجموعه مرجع



مشکل روش NMI



- در این مثال هر دو افراز فوق در خوشه‌بندی اولیه و همچنین در مجموعه مرجع مشاهده شده‌اند.
- تعداد کل افرازهای موجود در مجموعه مرجع=۴۰
- تعداد مشاهده خوشه بالا-چپ از افراز a=۴ (۱۰٪ موارد)
- تعداد مشاهده خوشه راست از افراز b=۴ (۱۰٪ موارد)

مشکل تقارن

راهکار پیشنهادی برای رفع مشکل

- روش ماکزیمم
- استفاده از معیار AMM
- اطلاعات متقابل نرمال اصلاح‌شده (ENMI)

روش ماکزیمم

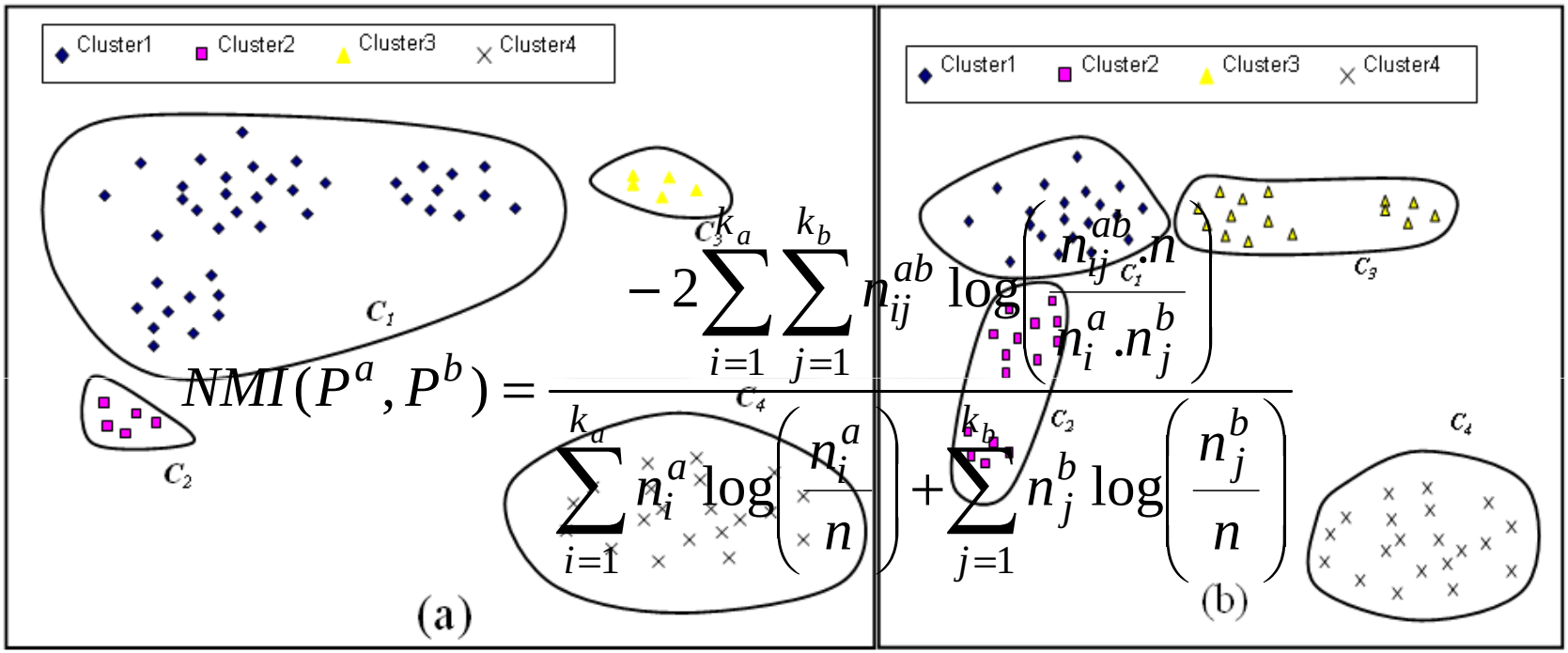
- داده‌های دو خوشه مکمل یکدیگر باشند
  - اجتماع داده‌های آنها شامل کل مجموعه داده شود
  - اشتراک داده‌های آنها تهی باشد

- تعداد خوشه‌های تشکیل‌دهنده مجموعه C^* در خوشه‌بندی مرجع عددی بزرگتر از یک باشد.

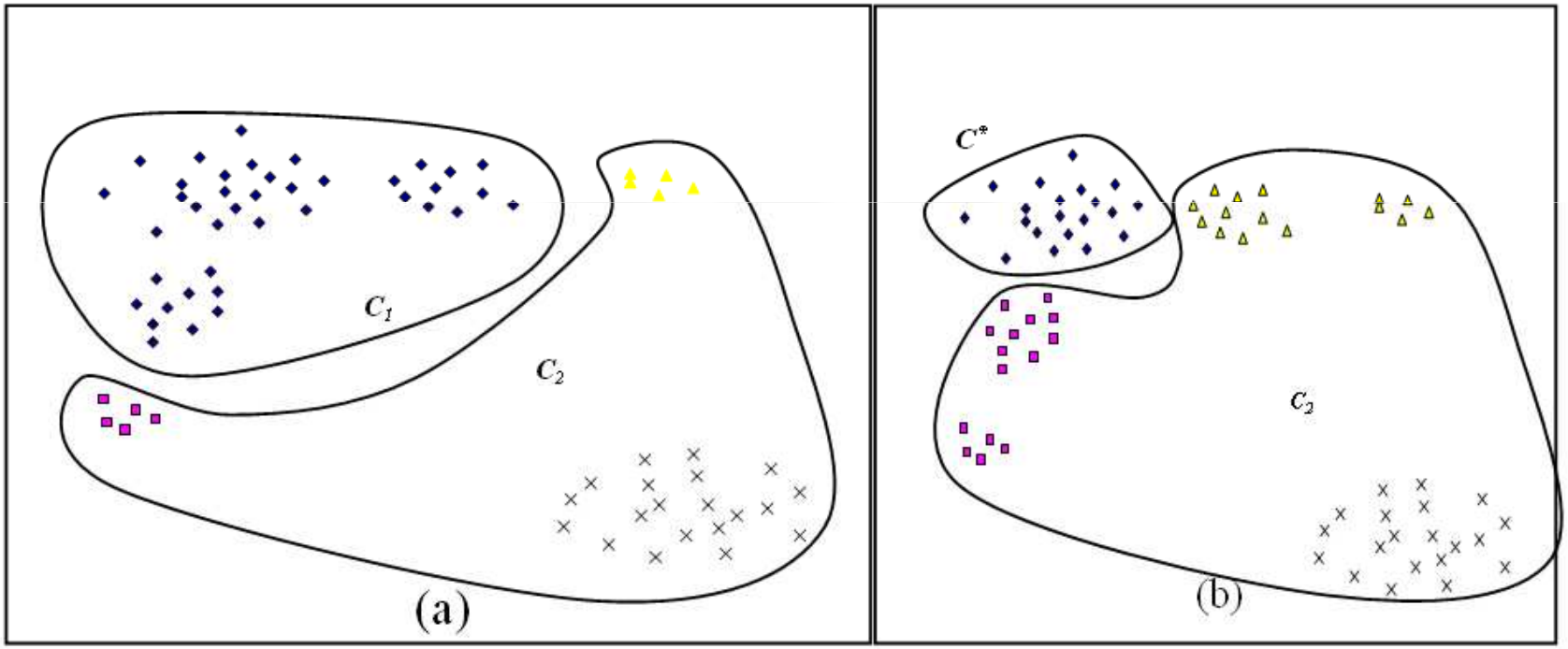
- C^* با ادغام دو یا بیشتر از خوشه‌ها به دست آید.

روش ماکزیمم:

انتخاب بزرگترین خوشه به عنوان مجموعه C^*، از بین تمام خوشه‌های موجود در مجموعه مرجع که شرط شباهت (بالای ۵۰٪ نمونه‌هایشان متعلق به خوشه C_1 باشد) را ارضا می‌کنند.



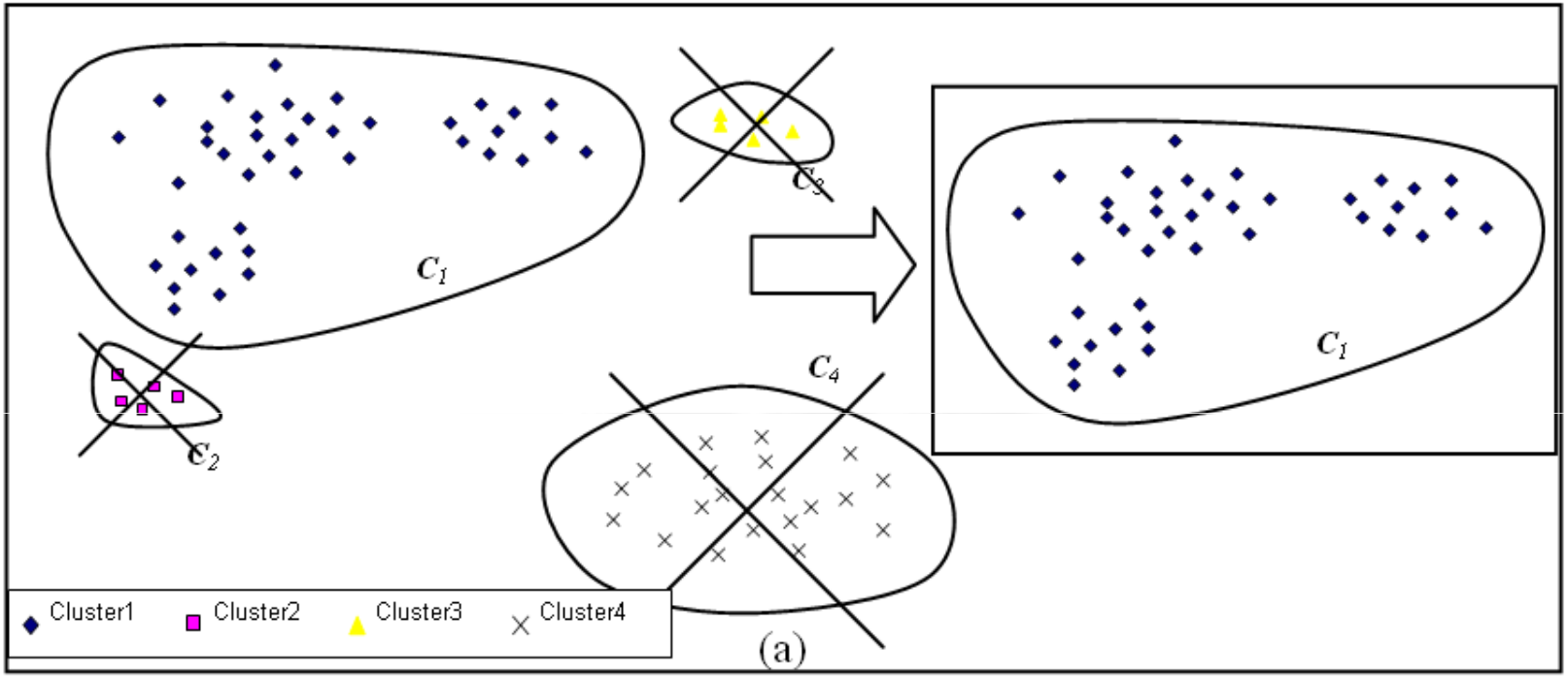
$$NMI(P^a, P^b) = \frac{-2\sum_{i=1}^{k_a}\sum_{j=1}^{k_b} n_{ij}^{ab} \log\left(\frac{n_{ij}^{ab}.n}{n_i^a.n_j^b}\right)}{\sum_{i=1}^{k_a} n_i^a \log\left(\frac{n_i^a}{n}\right) + \sum_{j=1}^{k_b} n_j^b \log\left(\frac{n_j^b}{n}\right)}$$

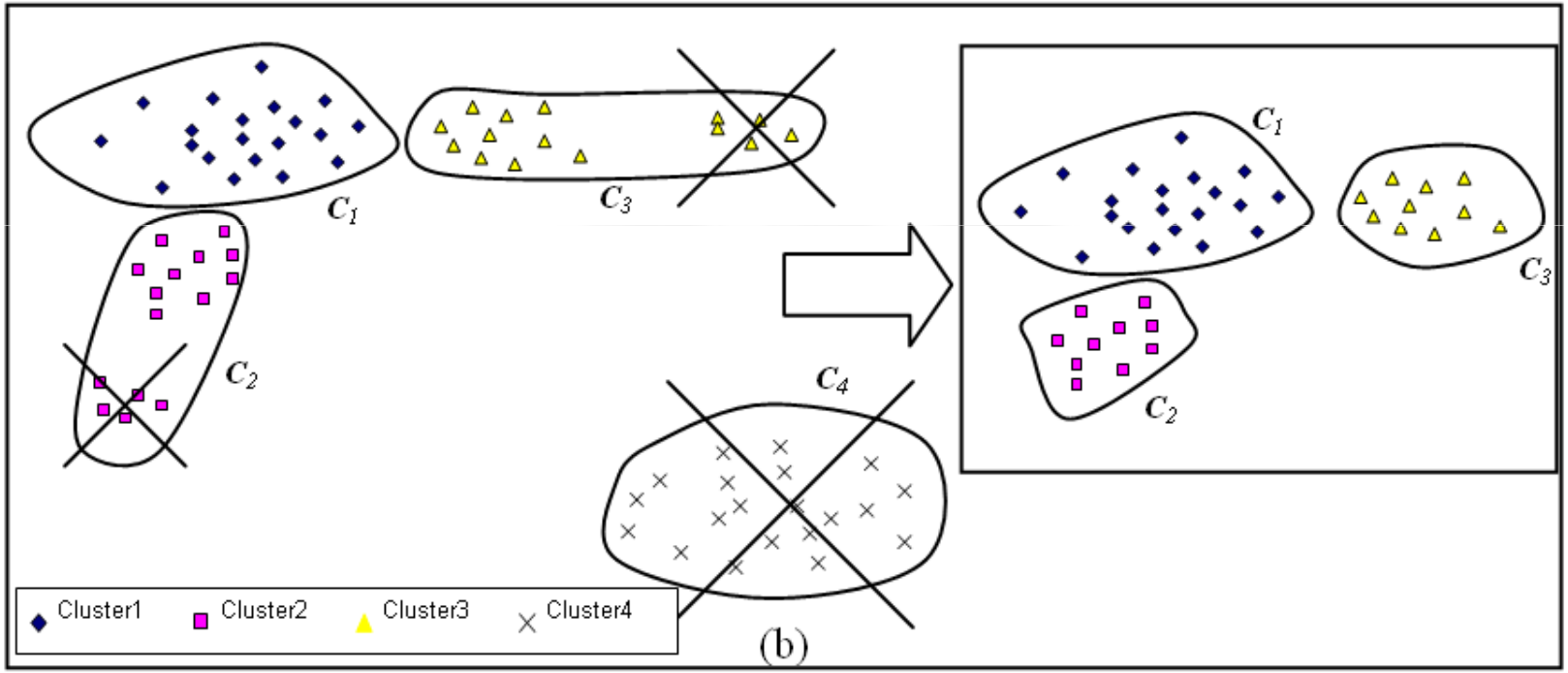


استفاده از معیار نامتقارن AMM

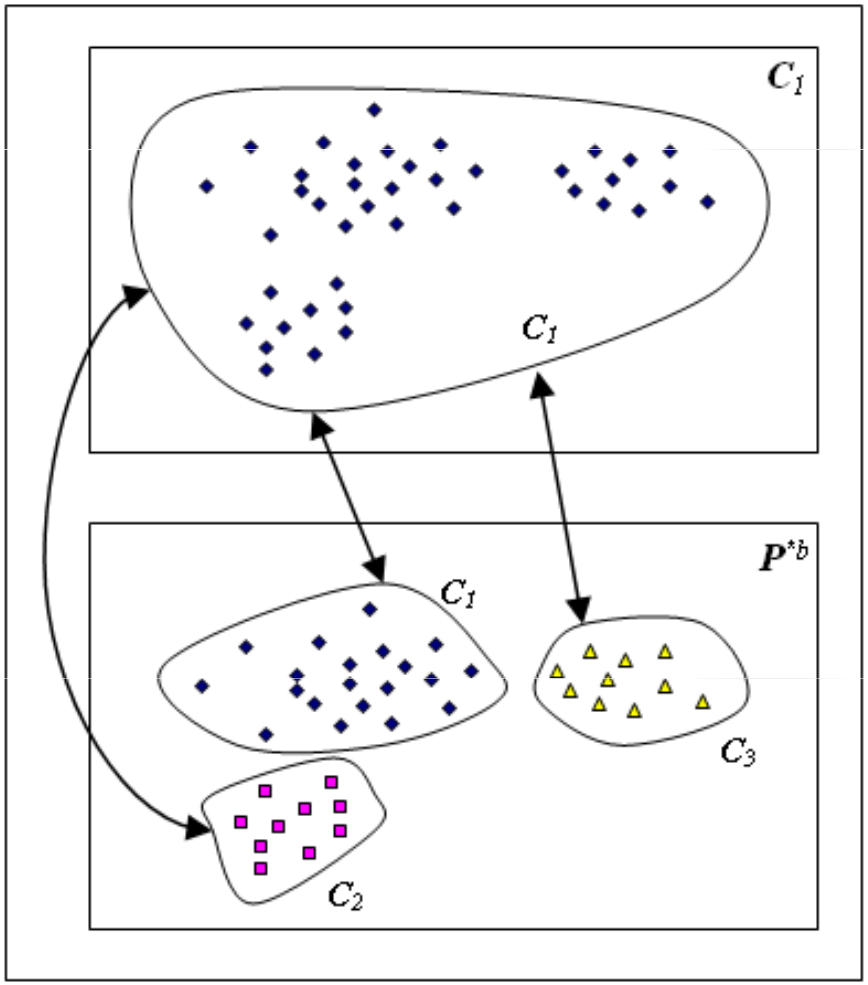
- از آن جایی که در این روش خوشه‌های غیر از خوشه مورد نظر در افراز a نادیده گرفته می‌شوند، این روش مشکل تقارن را نخواهد داشت.

- در این روش:
  - همه خوشه‌ها در افراز a به جز خوشه C_1 حذف می‌شوند.
  - همه خوشه‌ها در افراز b که شامل هیچ نمونه‌ی متناظری از خوشه C_1 نیستند، حذف می‌شوند.
  - خوشه‌های موجود در افراز b که شامل تعدادی از نمونه‌های متناظر خوشه C_1 هستند، از وجود دیگر نمونه‌ها پاک‌سازی می‌شوند.


C_1
C_2
C_3
C_4
C_1
Cluster1
Cluster2
Cluster3
Cluster4
(a)


C_1
C_2
C_3
C_4
C_1
C_2
C_3
Cluster1
Cluster2
Cluster3
Cluster4
(b)

معیار نامتقارن AMM



در این مرحله اطلاعات متقابل بین خوشه‌های P^a و P^b محاسبه می‌شود.

$$NMI(P^a, P^b) = \frac{-2\sum_{i=1}^{k_a}\sum_{j=1}^{k_b} n_{ij}^{ab} \log\left(\frac{n_{ij}^{ab}.n}{n_i^a.n_j^b}\right)}{\sum_{i=1}^{k_a} n_i^a \log\left(\frac{n_i^a}{n}\right) + \sum_{j=1}^{k_b} n_j^b \log\left(\frac{n_j^b}{n}\right)}$$

- در اینجا به جای تغییر مسئله به صورتی که بتوان از معیار NMI استفاده کرد، معیار به گونه‌ای تغییر داده خواهد شد تا بتوان اطلاعات متقابل نامتقارن را بین یک خوشه و یک افراز به دست آورد.

باقیمانده از n_i^a افراز a و n_j^{b*} محاسبه می‌شود.

$$AMM(C_i^a, P^{b*}) = \frac{-2\log\left(\frac{n}{n_i^a}\right)\sum_{j=1}^{k_{b*}} n_j^{b*}}{n_i^a \log\left(\frac{n_i^a}{n}\right) + \sum_{j=1}^{k_{b*}} n_j^{b*} \log\left(\frac{n_j^{b*}}{n}\right)}$$

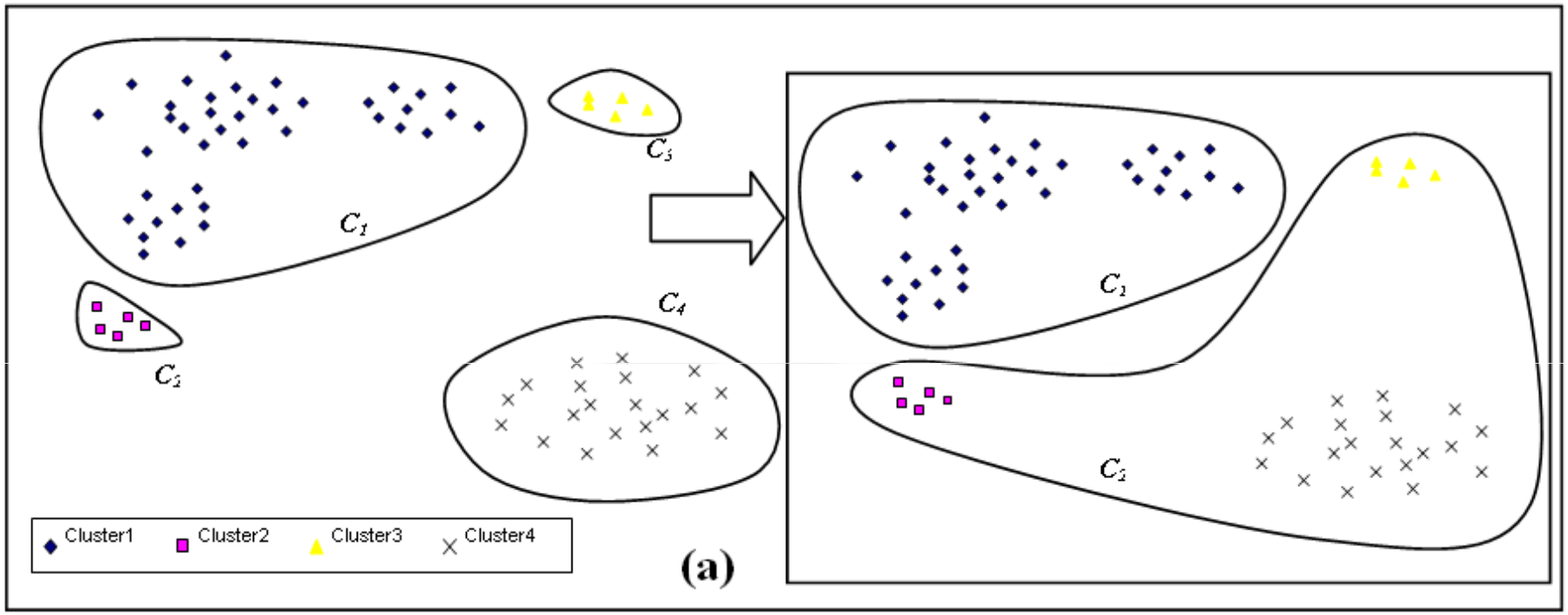
$$Stability(C_i) = AAMM(C_i) = \frac{1}{M}\sum_{j=1}^{M} AMM(C_i^a, P_j^{b*})$$

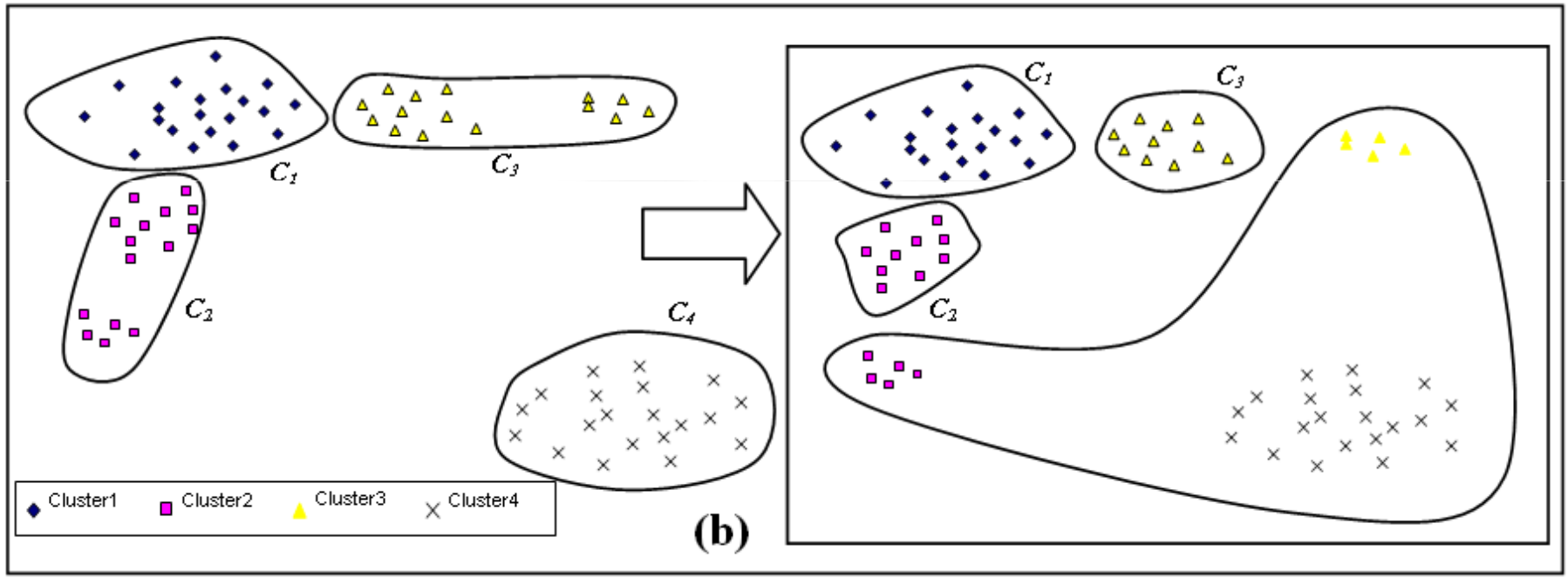
اطلاعات متقابل نرمال اصلاح شده (ENMI)

- مشابه روش AMM می‌باشد، با این تفاوت:

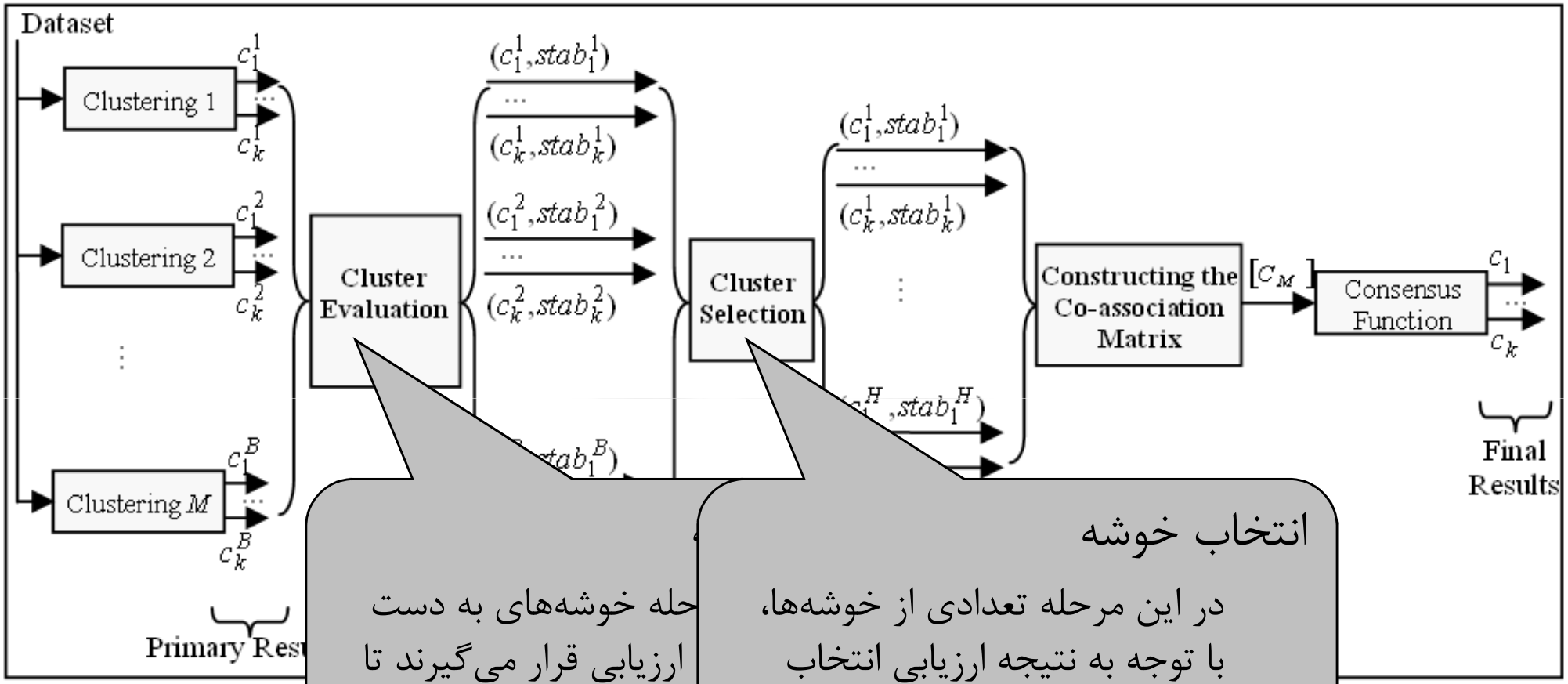
در روش AMM خوشه‌هایی از خوشه‌بندی مرجع که هیچ داده‌ی متناظری با C_1 نداشته باشند، در مرحله اول حذف می‌شوند. در صورتی که در این روش هیچ نمونه‌ای از داده‌ها حذف نمی‌شود.

در این روش، مجموعه تمام نمونه‌های غیرمتناظر با نمونه‌های خوشه C_1 به عنوان یک خوشه دیگر در محاسبه میزان پایداری شرکت می‌کنند.


C_1
C_2
C_3
C_4
Cluster1
Cluster2
Cluster3
Cluster4
(a)


C_1
C_2
C_3
C_4
Cluster1
Cluster2
Cluster3
Cluster4
(b)

چهارچوب کلی روش پیشنهادی



انتخاب خوشه

در این مرحله تعدادی از خوشه‌ها، با توجه به نتیجه ارزیابی انتخاب می‌شوند.

حله خوشه‌های به دست
ارزیابی قرار می‌گیرند تا
خوشه مشخص شود.

گام دوم: انتخاب زیرمجموعه‌ای از خوشه‌های اولیه

- در این مرحله عمل انتخاب خوشه‌ها با توجه به مقدار پایداری خوشه انجام می‌شود.

- روش‌های پیشنهادی:
  - اعمال آستانه
  - انتخاب تطبیقی
  - تنظیم پراکندگی و کیفیت

اعمال آستانه

- انتخاب خوشه‌های پایدارتر از خوشه‌بندی‌های اولیه با اعمال آستانه روی مقدار پایداری
- در این روش ماتریس همبستگی تنها از خوشه‌های پایدارتر تشکیل می‌شود.
- ترجیح دادن کیفیت به پراکندگی

انتخاب تطبیقی

- آیا همواره استفاده از خوشه‌های پایدارتر بهترین جواب ممکن هستند؟

- عمل انتخاب زیرمجموعه‌ای از خوشه‌ها در این روش با توجه به:

  مقدار پایداری خوشه‌ها و میزان سادگی مجموعه داده

  به صورت تطبیقی صورت می‌گیرد.

- روش کار برای انتخاب تطبیقی:
  - محاسبه میزان سادگی مجموعه داده
  - انتخاب زیرمجموعه‌ای مشخص از خوشه‌ها برای مجموعه داده مشخص

محاسبه میزان سادگی مجموعه داده

- برای محاسبه میزان سادگی مجموعه داده:
  - ابتدا معیار پایداری را برای یک افراز اولیه مستقل تعریف می‌کنیم.
  - سپس، سادگی مجموعه داده را مبتنی بر پایداری افرازهای اولیه تعریف می‌کنیم.

$$Stability(P) = \frac{1}{N}\sum_{i=1}^{k}|C_i|Stability(C_i)$$

تعداد خوشه‌ها

تعداد اعضای

پایداری یک افراز مستقل P

نمونه‌های افراز P

i-امین افراز اولیه

$$Simplicity(D) = \frac{1}{B}\sum_{i=1}^{B}Stability(P_i)$$

سادگی مجموعه داده D

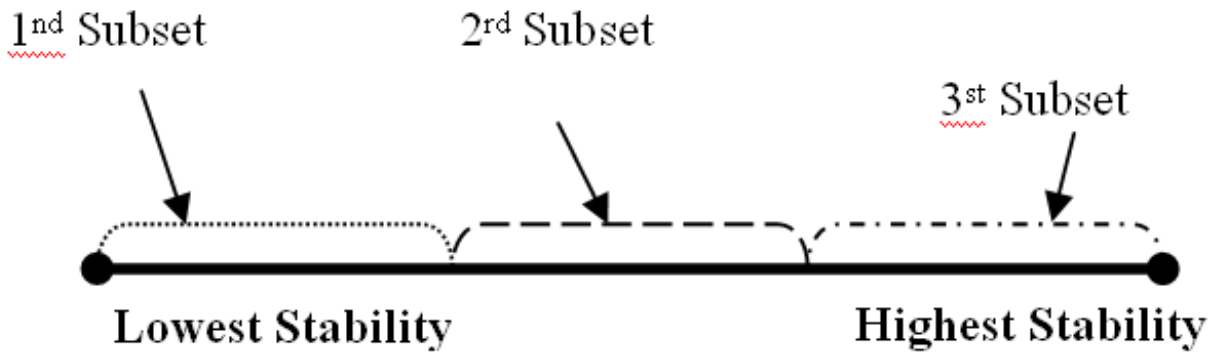
تعداد کل نتایج اولیه

رده‌بندی مجموعه داده‌ها

- هر چه این مقدار سادگی بیشترباشد، مجموعه داده راحت‌تر است.
- دسته‌بندی مجموعه داده‌ها:
  - سخت (میزان سادگی کمتر از ۰.۵)
  - متوسط (در بازه [۰.۵-۰.۵۵])
  - راحت (بالای ۰.۵۵)

$$f(\textit{Simplicity of Dataset}) = \textit{Special Subset of Clusters}$$

تخصیص زیرمجموعه مشخص از نتایج اولیه به هر رده



- تولید مجموعه داده مصنوعی (۱۵ مجموعه داده به ازای هر رده) و انجام آزمایش
- نتایج آزمایشات استفاده از زیرمجموعه سوم را برای هر سه رده پیشنهاد می‌کند.

تنظیم پراکندگی و کیفیت

- در نظر گرفتن همزمان هر دو معیار کیفیت و پراکندگی خوشه‌ها
- یافتن یک موازنه منطقی بین این دو معیار

- راهکارهای پیشنهادی برای تنظیم پراکندگی و کیفیت:
  - راهکار اول: مقایسه با شبیه‌ترین
  - راهکار دوم: مقایسه با یک همسایگی از شبیه‌ترین‌ها
  - راهکار سوم: خوشه‌بندی خوشه‌ها

راهکار اول: مقایسه با شبیه‌ترین

- ابتدا خوشه‌های اولیه بر اساس یک معیار رتبه‌بندی می‌شوند. (معیار AMM)
- در مرحله بعد به ترتیب با شروع از خوشه‌های بهتر، آنهایی انتخاب می‌شوند که میزان پراکندگی‌شان با خوشه‌های انتخاب شده، بیشتر از یک مقدار آستانه باشد.

```
S := {};            //Subset of Selected Clusters
T := Sort total primary clusters according to AMM;   //Sorted Total Primary Clusters
Initialization
    T := Sort total clusters according to AMM;
    S := {the most stable cluster};
For i := 2 to length (T)
    current := i-th cluster of T;
    similar := Find the most similar cluster in S with the current
    If distance(current,similar)>th
       Add current into S;
    End If;
End For;
Return S;
```

راهکار دوم: مقایسه با یک همسایگی از شبیه‌ترین‌ها

- ابتدا خوشه‌های اولیه بر اساس کیفیت رتبه‌دهی می‌شوند. (معیار AMM)
- به ترتیب از خوشه‌های بهتر شروع می‌کنیم و یک خوشه را بر می‌داریم.
- خوشه‌هایی از مجموعه S که بیشترین شباهت را به خوشه مورد نظر دارند
- در نهایت، اگر میانگین پراکندگی خوشه مورد نظر با خوشه‌های مجموعه شبیه‌ترین‌ها بیشتر از یک مقدار آستانه باشد، آن خوشه انتخاب می‌شود

```
S := {};                    //Subset of Selected Clusters
T                           //Sorted Total Primary Clusters
Initialization
    T := Sort total clusters according to AMM;
    S := {the most stable cluster};
For i := 2 to length (T)
current := i-th cluster of T;
For j := 1 to length(S)
    temp := j-th cluster of S;
    If similarity(current,temp)>th1
            Add temp into similar;
    End If;
End For;
If mean_distance(current,similar)>th2
    Add current into S;
End If;
End For;

Return S;
```

راهکار سوم: خوشه‌بندی خوشه‌ها

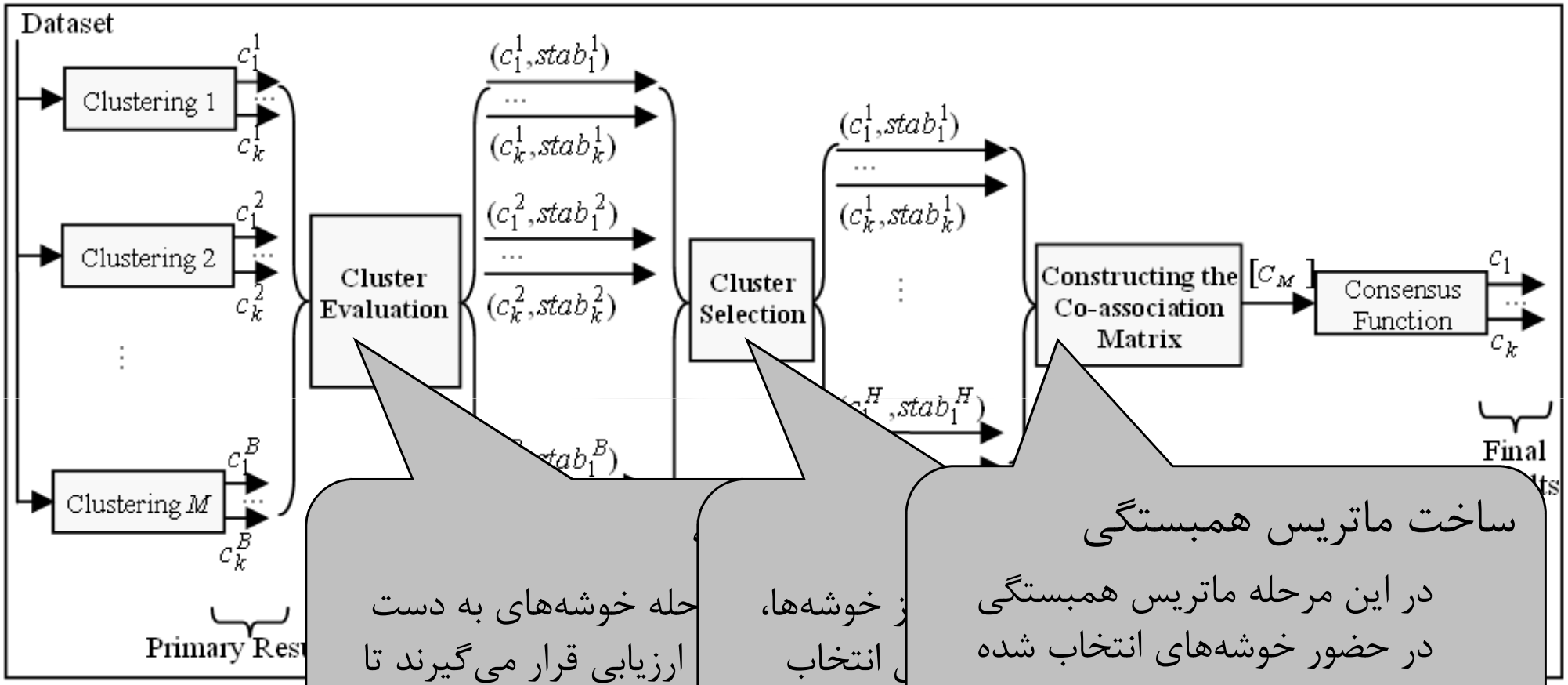
- اگرچه می‌توان با تنظیم پارامترها در دو راهکار اول و دوم به نتایج چشمگیری دست یافت، تنظیم پارامترهای مقادیر آستانه در این دو روش خود یکی از مشکلات این روش ها می‌باشد.

این روش هیچ پارامتری ندارد.

- در این روش:
  - یک خوشه‌بندی بر روی خوشه‌های اولیه صورت می‌گیرد.
    - تعداد خوشه‌های اعمال شده برای این خوشه‌بندی برابر است با ۳۳٪ تعداد کل خوشه‌های اولیه
  - در مرحله بعد از هر گروه پایدارترین خوشه انتخاب می‌شود.

```
S := {};                                   //Subset of Selected Clusters
G := Apply a clustering technique over all primary clusters
For i := 1 to length(G)
    current := i-th group of G;
    If size(current)>1
        Add the most stable cluster of in the current into S;
    End If;
End For;
Return S;
```

چهارچوب کلی روش پیشنهادی



گام سوم: ساخت ماتریس همبستگی از زیرمجموعه‌ای از خوشه‌ها

- در این مرحله خوشه‌های انتخاب شده با هم ترکیب شده و خوشه‌های نهایی از آنها به دست می‌آید.
- خوشه‌بندی انباشت مدارک (EAC)، توسط فرد و جین (۲۰۰۵)

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

- $n_{i,j}$ تعداد دفعاتی است که جفت نمونه‌های i و j با هم در یک خوشه گروه‌بندی شده‌اند.
- $m_{i,j}$ تعداد نمونه‌برداری‌هایی است که هر دوی این جفت نمونه‌ها به طور همزمان در آن ظاهر شده‌اند.

خوشه‌بندی انباشت مدارک توسعه یافته (EEAC)

$$C(i, j) = \frac{n_{i,j}}{\max(n_i, n_j)}$$

- $n_{i,j}$ تعداد دفعاتی است که جفت نمونه‌های i و j با هم در یک خوشه گروه‌بندی شده‌اند.
- n_i تعداد دفعاتی است که نمونه i در خوشه‌های انتخاب شده ظاهر شده است.

روش اشتراک به اجتماع (ItoU)

- مشابه روش انباشت مدارک توسعه یافته
- ایده اصلی در این روش شمارش تمام حالت‌های ممکن نمونه‌های i و j نسبت به هم

$$C(i,j) = \frac{\bigcap(n_i, n_j)}{\bigcup(n_i, n_j)} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}}$$

- مقدمه‌ای بر خوشه‌بندی ترکیبی
- روش پیشنهادی
  - ارزیابی خوشه
  - انتخاب خوشه
  - ساخت ماتریس همبستگی
- **نتایج آزمایشات**
- جمع‌بندی و کارهای آینده

مجموعه داده‌ها

- ۱۰ مجموعه داده استاندارد (UCI)

خلاصه‌ای از مشخصه‌های مجموعه داده‌های استاندارد مورد استفاده

	Class	Features	Samples
Glass	6	9	214
Breast-Cancer	2	9	683
Wine	3	13	178
Bupa	2	6	345
Yeast	10	8	1484
Iris	3	4	150
SAHeart	2	9	462
Ionosphere	2	34	351
Halfrings	2	2	400
Galaxy	7	4	323

پارامترهای مورد استفاده

- K-means به عنوان الگوریتم پایه
  - مقادیر اولیه تصادفی (Seed Points)
  - نمونه‌برداری ۵۰٪ از داده‌ها
  - تعداد خوشه‌ها با اندازه‌های k، k+1، k+2 و k+3

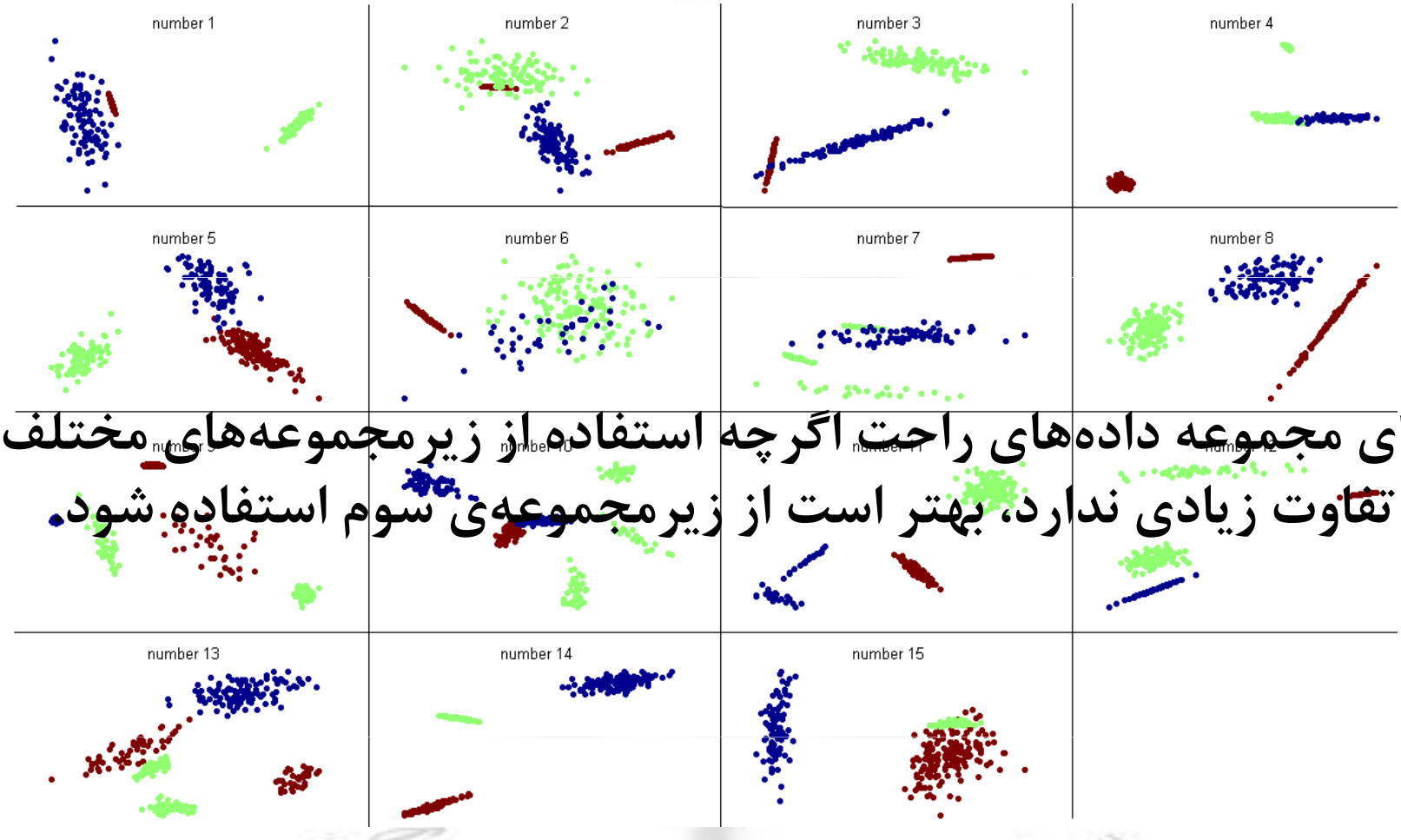
- تعداد نتایج اولیه تولید شده در تمام روش‌ها ثابت و برابر با ۱۲۰
- استخراج نتایج نهایی از ماتریس همبستگی: روش اتصال منفرد (Single Linkage)

نتایج آزمایشات

روش ارزیابی خوشه	روش ساخت ماتریس همبستگی	مجموعه داده‌های استاندارد										
		N. Breast Cancer	Iris	N. Bupa	N. SAHeart	Ionosphere	N. Glass	Halfrings	N. Galaxy	N. Yeast	Wine	N. Wine
NMI	ItoU	95.02	88.67	54.78	63.42	70.09	44.86	74.50	29.41	42.86	70.22	96.63
	EEAC	95.73	76.13	54.33	63.36	70.60	**47.76**	74.48	31.27	42.93	69.38	85.17
MAX	ItoU	96.93	**90.00**	54.78	**64.50**	**71.51**	44.86	87.25	29.41	48.45	71.35	97.75
	EEAC	96.49	84.87	**57.42**	63.87	57.75	44.35	74.55	29.85	**51.27**	70.00	94.44
AMM	ItoU	95.43	88.00	54.73	63.42	**71.51**	44.39	74.50	29.69	48.52	70.73	96.63
	EEAC	95.46	**90.00**	55.07	63.85	70.66	45.79	54.00	30.65	53.10	70.23	96.63
ENMI	ItoU	96.78	**90.00**	55.07	**64.50**	**71.51**	45.79	**88.25**	30.03	50.47	70.23	**98.32**
	EEAC	96.93	88.67	54.78	63.20	71.23	43.93	**88.00**	30.65	50.47	70.23	97.19
D&Q	1	97.66	97.33	55.36	68.83	72.93	50.47	87.25	35.29	56.67	72.47	98.31
	2	97.07	91.33	55.36	68.02	74.36	53.74	76.50	33.44	54.33	71.35	98.31
	3	**97.05**	**90.00**	55.00	63.83	70.91	47.20	83.25	**31.36**	50.18	**71.42**	97.75
Adaptive		95.43	88.00	54.73	63.42	**71.51**	44.39	74.50	29.69	48.52	70.73	96.63
EAC (Full Ens.)		95.17	89.33	54.49	63.20	70.66	46.26	74.50	30.96	44.21	70.22	96.63
Azimi		96.91	89.33	54.75	56.06	70.74	45.05	67.70	29.97	43.40	60.95	96.63

آزمایشات روی روش تطبیقی

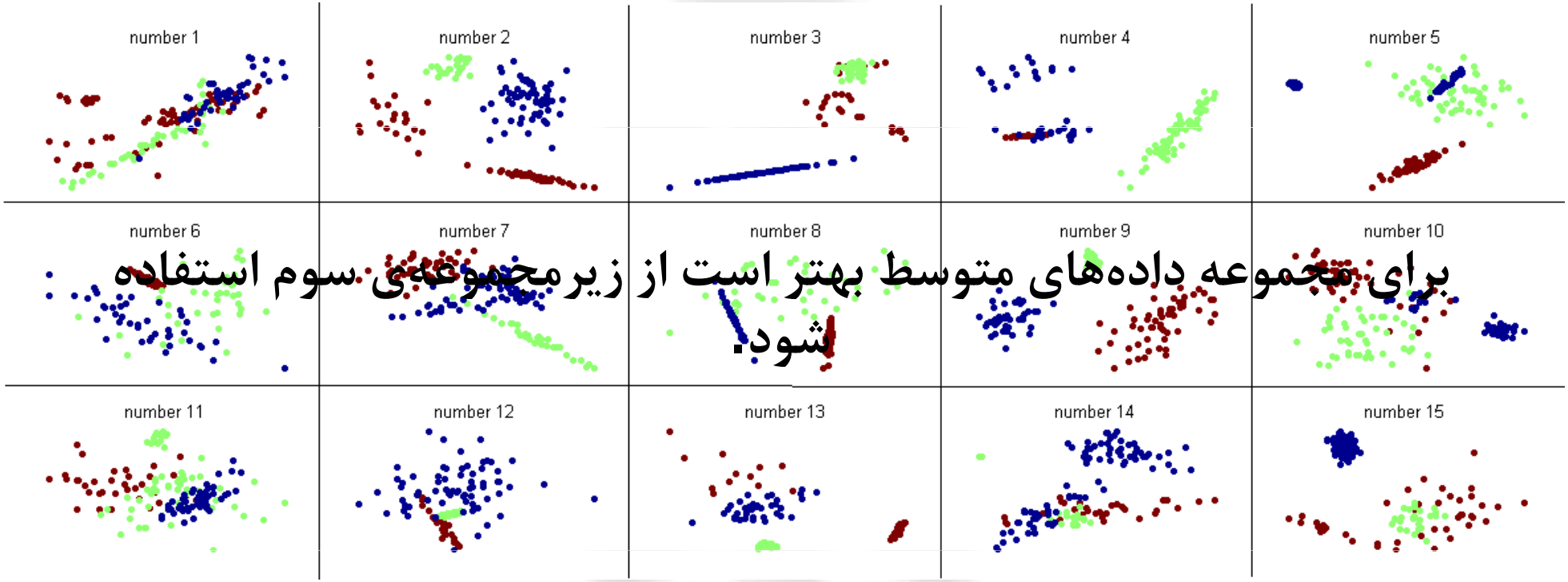
- یافتن رابطه‌ای معنی‌دار بین میزان سادگی مجموعه داده و زیرمجموعه‌ای از خوشه‌ها
- تولید خودکار ۴۵ مجموعه داده:
  - ۱۵ مجموعه داده راحت
  - ۱۵ مجموعه داده متوسط
  - ۱۵ مجموعه داده سخت
- خصوصیات مجموعه داده‌های تولید شده:
  - ۳۰۰ داده
  - ۳ رده
  - ۲ بعدی
  - توزیع داده‌ها به صورت تصادفی با یکی از توزیع‌های زیر حول مرکز خوشه:
    - نرمال گوسین
    - توزیع k
    - توزیع یکنواخت در یک (تولید خوشه میله‌ای)
    - توزیع یکنواخت در دو بعد



برای مجموعه داده‌های راحت اگرچه استفاده از زیرمجموعه‌های مختلف تفاوت زیادی ندارد، بهتر است از زیرمجموعه‌ی سوم استفاده شود.

مجموعه داده	میزان سادگی	نرخ تشخیص با استفاده از زیرمجموعه اول	نرخ تشخیص با استفاده از زیرمجموعه دوم	نرخ تشخیص با استفاده از زیرمجموعه سوم
راحت	0.59	79.63	79.14	**79.86**

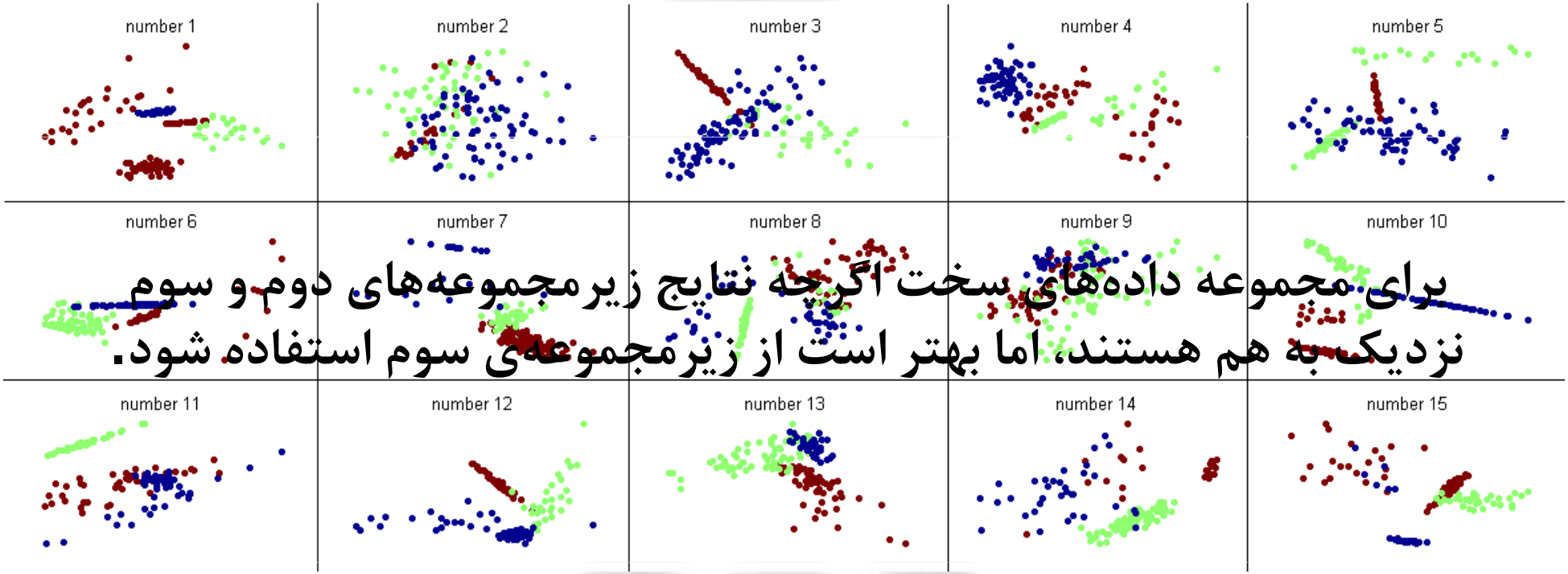
مجموعه داده متوسط



برای مجموعه داده‌های متوسط بهتر است از زیرمجموعه‌ی سوم استفاده شود.

مجموعه داده	میزان سادگی	نرخ تشخیص با استفاده از زیرمجموعه اول	نرخ تشخیص با استفاده از زیرمجموعه دوم	نرخ تشخیص با استفاده از زیرمجموعه سوم
متوسط	0.52	69.23	74.29	**77.51**

مجموعه داده سخت



برای مجموعه داده‌های سخت اگرچه نتایج زیرمجموعه‌های دوم و سوم نزدیک به هم هستند، اما بهتر است از زیرمجموعه‌ی سوم استفاده شود.

مجموعه داده	میزان سادگی	نرخ تشخیص با استفاده از زیرمجموعه اول	نرخ تشخیص با استفاده از زیرمجموعه دوم	نرخ تشخیص با استفاده از زیرمجموعه سوم
سخت	0.48	59.40	68.36	**68.58**

روش تطبیقی در یک نگاه

مجموعه داده	میزان سادگی	نرخ تشخیص با استفاده از زیرمجموعه اول	نرخ تشخیص با استفاده از زیرمجموعه دوم	نرخ تشخیص با استفاده از زیرمجموعه سوم
راحت	0.59	79.63	79.14	**79.86**
متوسط	0.52	69.23	74.29	**77.51**
سخت	0.48	59.40	68.36	**68.58**

- نتایج نشان می‌دهند که همواره استفاده از زیرمجموعه‌ی سوم نتیجه بهتری می‌دهد.
- از آنجایی که میزان سادگی مجموعه داده نقشی در تعیین زیرمجموعه‌ی مناسب ندارد، بنابراین نتیجه استفاده از روش تطبیقی مشابه با روش اعمال آستانه خواهد بود.

روش ارزیابی خوشه	روش ساخت ماتریس همبستگی	مجموعه داده‌های استاندارد										
		N. Breast Cancer	Iris	N. Bupa	N. SAHeart	Ionosphere	N. Glass	Halfrings	N. Galaxy	N. Yeast	Wine	N. Wine
NMI	ItoU	95.02	88.67	54.78	63.42	70.09	44.86	74.50	29.41	42.86	70.22	96.63
	EEAC	95.73	76.13	54.33	63.36	70.60	**47.76**	74.48	31.27	42.93	69.38	85.17
MAX	ItoU	96.93	**90.00**	54.78	**64.50**	**71.51**	44.86	87.25	29.41	48.45	71.35	97.75
	EEAC	96.49	84.87	**57.42**	63.87	57.75	44.35	74.55	29.85	**51.27**	70.00	94.44
AMM	ItoU	95.43	88.00	54.73	63.42	**71.51**	44.39	74.50	29.69	48.52	70.73	96.63
	EEAC	95.46	**90.00**	55.07	63.85	70.66	45.79	54.00	30.65	53.10	70.23	96.63
ENMI	ItoU	96.78	**90.00**	55.07	**64.50**	**71.51**	45.79	**88.25**	30.03	50.47	70.23	**98.32**
	EEAC	96.93	88.67	54.78	63.20	71.23	43.93	**88.00**	30.65	50.47	70.23	97.19
D&Q	1	97.66	97.33	55.36	68.83	72.93	50.47	87.25	35.29	56.67	72.47	98.31
	2	97.07	91.33	55.36	68.02	74.36	53.74	76.50	33.44	54.33	71.35	98.31
	3	**97.05**	**90.00**	55.00	63.83	70.91	47.20	83.25	**31.36**	50.18	**71.42**	97.75
Adaptive		95.43	88.00	54.73	63.42	**71.51**	44.39	74.50	29.69	48.52	70.73	96.63
EAC (Full Ens.)		95.17	89.33	54.49	63.20	70.66	46.26	74.50	30.96	44.21	70.22	96.63
Azimi		96.91	89.33	54.75	56.06	70.74	45.05	67.70	29.97	43.40	60.95	96.63

- ارزیابی خوشه‌های اولیه و تعیین میزان سادگی مجموعه داده با معیار AMM
- ساخت ماتریس همبستگی با روش اشتراک به اجتماع

- مقدمه‌ای بر خوشه‌بندی ترکیبی
- روش پیشنهادی
  - ارزیابی خوشه
  - انتخاب خوشه
  - ساخت ماتریس همبستگی
- نتایج آزمایشات
- **جمع‌بندی و کارهای آینده**

جمع بندی

- در این پایان‌نامه چند روش مبتنی بر انتخاب زیرمجموعه‌ای از نتایج اولیه، برای بهبود کارایی خوشه‌بندی ترکیبی پیشنهاد شد.
- چهارچوب روش پیشنهادی:
  - ارزیابی خوشه
    - روش NMI
    - روش MAX
    - روش AMM
    - روش ENMI
  - انتخاب زیرمجموعه‌ای از خوشه‌ها
    - روش اعمال آستانه
    - روش تطبیقی
    - روش تنظیم پراکندگی و کیفیت
  - ساخت ماتریس همبستگی
    - انباشت مدارک توسعه یافته
    - اشتراک به اجتماع

نتایج نشان می‌دهند که اگرچه در روش‌های پیشنهادی تنها ۳۳٪ از نتایج اولیه در ترکیب نهایی استفاده می‌شوند، می‌توانند کارایی خوشه‌بندی ترکیبی را حتی نسبت به روش ترکیب کامل هم بهتر کنند.

کارهای آینده

- ارایه راهکاری خودکار برای تعیین بهینه مقدار آستانه در روش‌های پیشنهادی
- استفاده از حالت‌های (زیرمجموعه‌های) بیشتر و پیچیده‌تر در روش تطبیقی
- تعیین خودکار پارامترها در راهکارهای اول و دوم از روش تنظیم پراکندگی و کیفیت
- جستجوی بیشتر برای ارایه معیارهای مناسب برای ارزیابی یک خوشه و یا یک افراز
- الگوریتم‌های تکاملی نظیر الگوریتم ژنتیک

تعدادی از مقالات (۱)

- ***H. Alizadeh*** *and B. Minaei-Bidgoli, "**Improving the Performance of Clustering Ensemble Using Cluster Selection**", Journal of Zhejiang University SCIENCE-A (JZUS), (Thomson ISI Indexed), ISSN: 1673-565X (under review).*
- *B. Minaei-Bidgoli, **H. Alizadeh**, H. Parvin and W. F. Punch, "**Effects of Resampling Method and Adaptation on Clustering Ensemble Efficacy**", International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), (Thomson ISI Indexed), ISSN: 0218-0014 (under review).*
- *Parvin H., **Alizadeh H.** and Minaei-Bidgoli B. (2009), **A New Method for Constructing Classifier Ensembles**, International Journal of Digital Content: Technology and its Application, JDCTA, ISSN: 1975-9339, (in press).*

تعدادی از مقالات (۲)

- ***H. Alizadeh***, *S.K. Amirgholipour, N.R. Seyedaghaee and B. Minaei-Bidgoli,* ***Using Clustering Ensemble in Classification Problems****, Fourth International Conference on Intelligent Computing and Information Systems (ICICIS09), Egypt, March 19-22, 2009 (Indexed by ACM Library).*
- *H. Parvin,* ***H. Alizadeh*** *and B. Minaei-Bidgoli,* ***Using Clustering for Generating Diversity in Classifier Ensemble****, International Journal of Digital Content: Technology and its Application, JDCTA, ISSN: 1975-9339, 2009 (in press).*
- ***Alizadeh H.,*** *Amirgholipour S.K., Seyedaghaee N.R. and Minaei-Bidgoli B. (2009),* ***Nearest Cluster Ensemble (NCE): Clustering Ensemble Based Approach for Improving the performance of K-Nearest Neighbor Algorithm****, 11th Conf. of the International Federation of Classification Societies, IFCS09, March 13–18. (in press).*
- *H. Parvin, M. Parvari,* ***H. Alizadeh*** *and B. Minaei-Bidgoli, "****Clustering Based Classifier Ensembles****", Fourth International Conference on Intelligent Computing and Information Systems (ICICIS09), Egypt, March 19-22, 2009 (Indexed by ACM Library).*

تعدادی از مقالات (۳)

- *M. Mohammadi, **H. Alizadeh** and B. Minaei-Bidgoli, "**Neural Network Ensembles using Clustering Ensemble and Genetic Algorithm**", 2008 3rd Int. Conf. on Convergence and hybrid Information Technology, ICCIT08, Nov. 11-13, 2008, pp.761-766, ISBN: 978-0-7695-3407-7, Published by IEEE CS.*

- ***علیزاده ح.***، مینایی بیدگلی ب.، (۱۳۸۷)، **خوشه‌بندی ترکیبی مبتنی بر زیرمجموعه‌ای از خوشه‌های اولیه**، مجله دانشکده فنی (JFE)، دانشگاه تهران (تحت بررسی).
- ***علیزاده ح.***، مینایی بیدگلی ب.، (۱۳۸۷)، **یک روش جدید برای خوشه‌بندی ترکیبی با استفاده از خوشه‌های پایدار**، پنجمین کنفرانس بین‌المللی مدیریت فناوری اطلاعات و ارتباطات، تهران، ایران.
- ***علیزاده ح.***، مینایی بیدگلی ب.، (۱۳۸۷)، ***ایجاد پراکندگی در خوشه‌بندی ترکیبی***، کنفرانس ملی مهندسی نرم‌افزار و کاربردهای آن، لاهیجان، ایران.
- ***علیزاده ح.***، مینایی بیدگلی ب.، (۱۳۸۷)، ***بررسی روش‌های ارزیابی خوشه‌بندی***، کنفرانس ملی مهندسی نرم‌افزار و کاربردهای آن، لاهیجان، ایران.

رشته تسبیح گر بگسست معذورم بدار

دستم اندر ساعد ساقی سیمین ساق بود

متشکرم