خوشه‌بندی انعطاف‌پذیر مبتنی بر چگالی داده‌های عظیم شبکه

مقدّمه

- هدف : خوشه‌بندی داده‌های عظیم
- چالش‌های پردازش داده‌های عظیم
  - محدودیت ذخیره‌سازی
  - محدودیت پردازشی
  - خرابی گره‌ها
- راهکارها
  - توسعهٔ روش‌های کلاسیک برای سکّوهای پردازش داده‌های عظیم

خوشه‌بندی

- مفهوم کلّی خوشه‌بندی
- شباهت نمونه‌های یک خوشه به هم
- تفاوت نمونه‌های خوشه‌های مختلف

دسته‌بندی روش‌های خوشه‌بندی

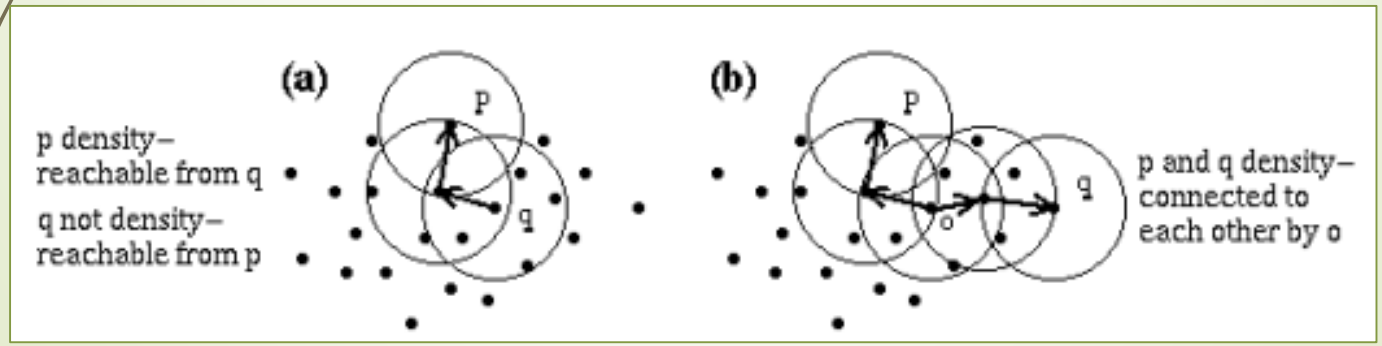
جزءبندی	مبتنی بر چگالی	سلسله‌مراتبی		روش‌های دیگر		
KMeans	DBSCAN	Agglomerative	Divisive	Subscpace	Evolutionary	...

خوشه‌بندی مبتنی بر چگالی

- مزایای خوشه‌بندی مبتنی بر چگالی جهت خوشه‌بندی داده‌های شبکه
  - توانایی شناسایی خوشه‌هایی با هر شکل
  - عدم نیاز به آگاهی از تعداد خوشه‌ها
  - برتری کیفی بر روش‌های دیگر در خوشه‌بندی این نوع داده

DBSCAN

- الگوریتم DBSCAN
  - بررسی نمونه‌ها در یک شعاع محدود (پارامتر eps)
  - شناسایی نمونه‌هایی با حداقل چگالی (پارامتر minPts) و گسترش خوشه‌ها



(Ester, Kriegel et al. 1996)

DBSCAN(ادامه)

- مزایا
  - امکان پیدا کردن خوشه‌هایی با هر توزیع
  - عدم نیاز به آگاهی از تعداد خوشه‌ها
  - امکان شناسایی نویز
- معایب
  - امکان ادغام خوشه‌های نزدیک به هم
  - عدم امکان تشخیص خوشه‌های دارای چگالی مختلف به صورت همزمان
  - نیاز به تخمین چگالی خوشه‌ها برای تنظیم پارامترها

توسعهٔ DBSCAN برای داده‌های عظیم

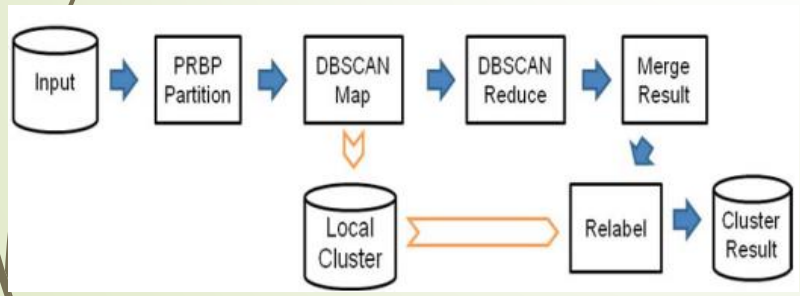
- چارچوب مورد استفاده

۱. تقسیم فضا به نواحی دارای هم‌پوشانی (شکل راست)

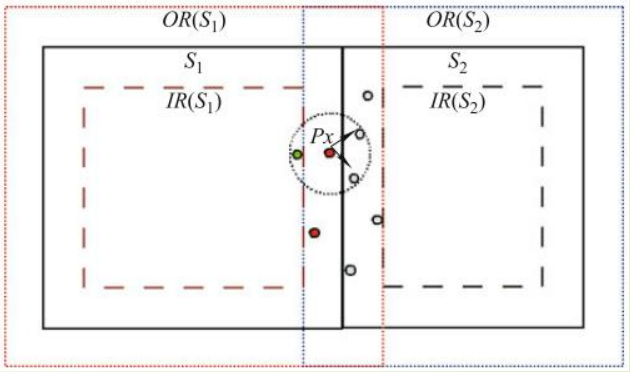
۲. خوشه‌بندی داده‌های هر ناحیه به صورت مجزّا

۳. ادغام نتایج نواحی مختلف با توجّه به نقاط موجود در هم‌پوشانی‌ها

- هدف: رسیدن به نتیجهٔ روش کلاسیک برای داده‌های عظیم



(Bi-Ru and Lin 2012)



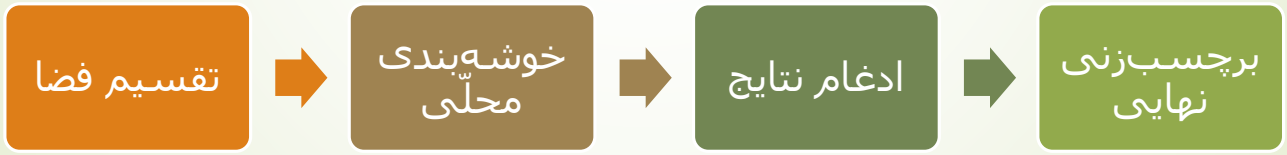
(He, Tan et al. 2014)

الگوریتم :

- روش خوشه‌بندی
  - خوشه‌بندی مبتنی بر چگالی (تمرکز بر خوشه‌هایی با چگالی متفاوت) KNNCA
- چارچوب پردازش توزیع‌شده
  - نگاشت‌کاهش
- رویکرد انتخابی
  - تقسیم فضا و خوشه‌بندی محلّی

روش کار (ادامه)

- تقسیم فضا جهت ممکن کردن پردازش داده‌ها در یک گره
- خوشه‌بندی داده‌های هر قسمت در یک گره
- ادغام نتایج خوشه‌های قسمت‌های مختلف



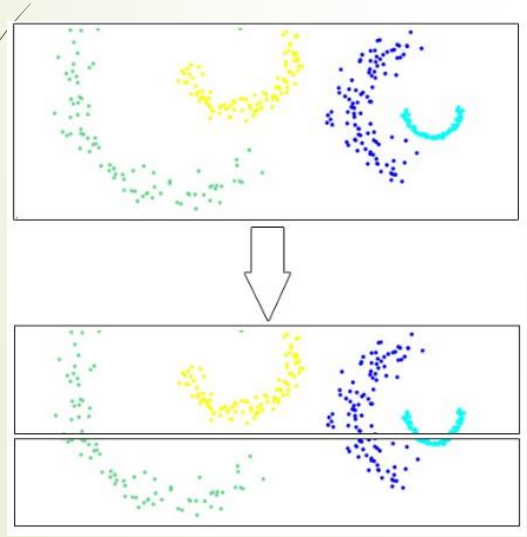
تقسیم فضای داده

- تقسیم فضا جهت حفظ همسایگی
- شروع تقسیم با کل فضا
- شکستن قسمت‌هایی با جمعیت بیشتر از max_samples در هر گام
- تصمیم‌گیری برای شکستن یک قسمت
  - انتخاب محور شکست
  - انتخاب نقطهٔ شکست

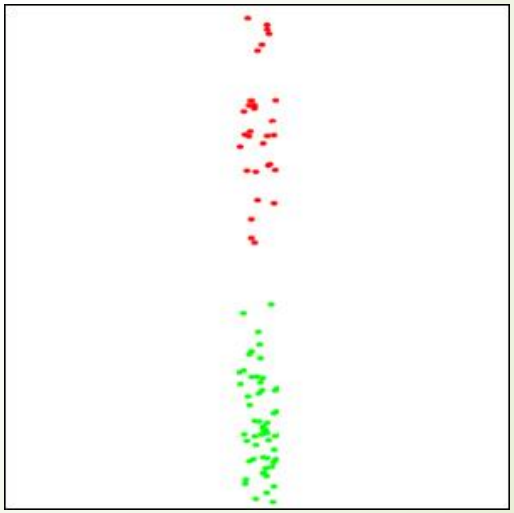
انتخاب نقطهٔ شکست	انتخاب محور شکست		
		واریانس نمونه‌ها	طول محور
	وسط محور	VM	LM
	میانگین	VA	LA

تقسیم فضای داده (ادامه)

- ملاحظات انتخاب محور
  - توزیع نمونه‌ها در زیرقسمت‌های حاصل
  - تعداد نمونه‌های مرزی حاصل



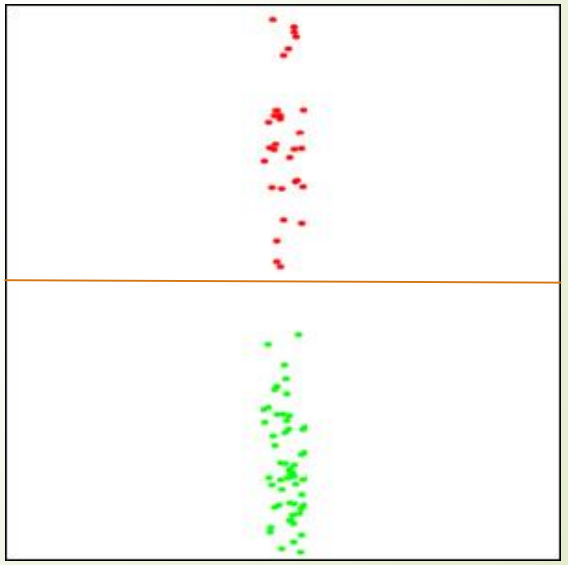
طول محور



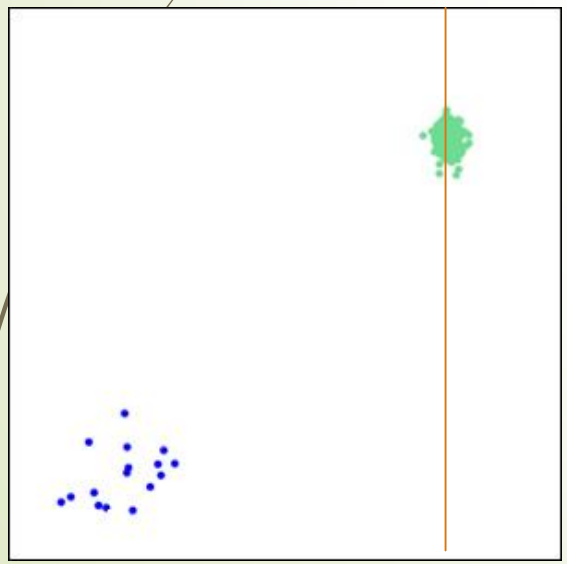
واریانس نمونه‌ها
در طول محور

تقسیم فضای داده (ادامه)

- ملاحظات انتخاب نقطهٔ شکست
  - توزیع متناسب نمونه‌ها در زیرقسمت‌ها
  - تعداد نمونه‌های مرزی حاصل



وسط محور



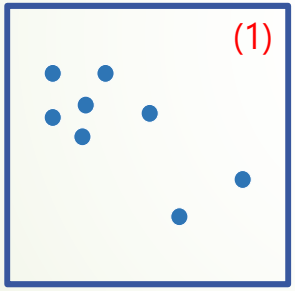
میانگین نمونه‌ها

خوشه‌بندی محلّی KNNCA

- تمرکز بر رویکرد **تفکّر انسانی** در هوش مصنوعی
- **رفع محدودیت** شعاع گسترش و همسایگی نمونه‌ها
- تخمین چگالی نمونه‌ها به صورت محلّی
- گسترش خوشه‌ها با توجّه به **اختلاف چگالی**

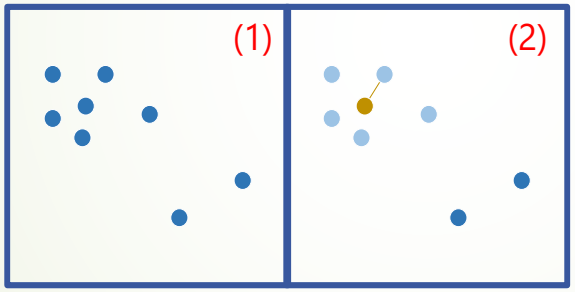
خوشه‌بندی محلّی KNNCA (ادامه)

- بررسی k همسایه بدون محدودیت در شعاع همسایگی



خوشه‌بندی محلّی KNNCA (ادامه)

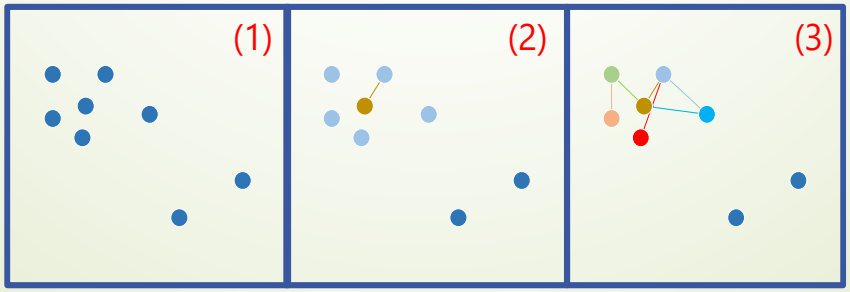
- چگالی تخمینی هر نمونه
  - استفاده از مفاهیمی مثل میانگین یا میانهٔ فاصله تا همسایه‌ها



خوشه‌بندی محلّی KNNCA (ادامه)

- گسترش خوشه با توجّه به اختلاف چگالی

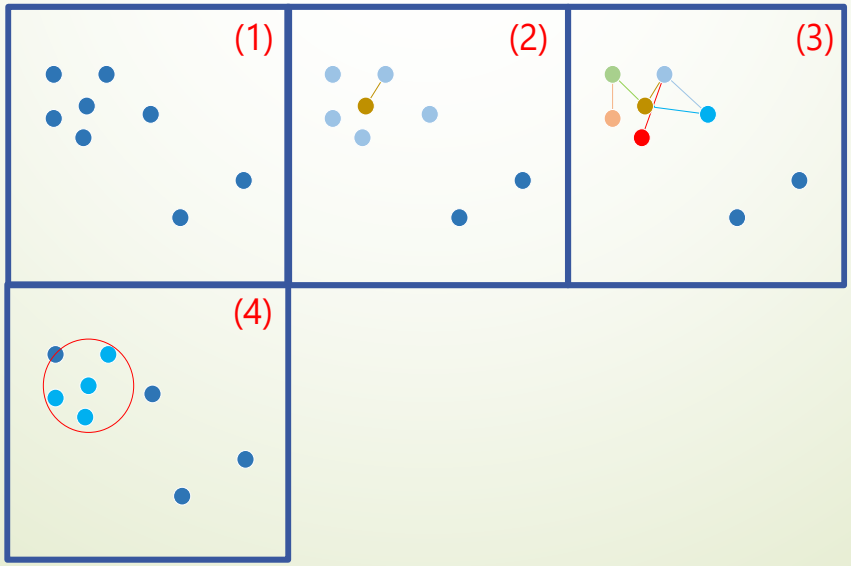
$$\frac{\left\| Density(i) - \frac{\sum_{j\ in\ knnbrs(i)} density(j)}{k} \right\|}{\max\left(Density(i), \frac{\sum_{j\ in\ knnbrs(i)} density(j)}{k}\right)} < f \quad (3)$$



خوشه‌بندی محلّی KNNCA (ادامه)

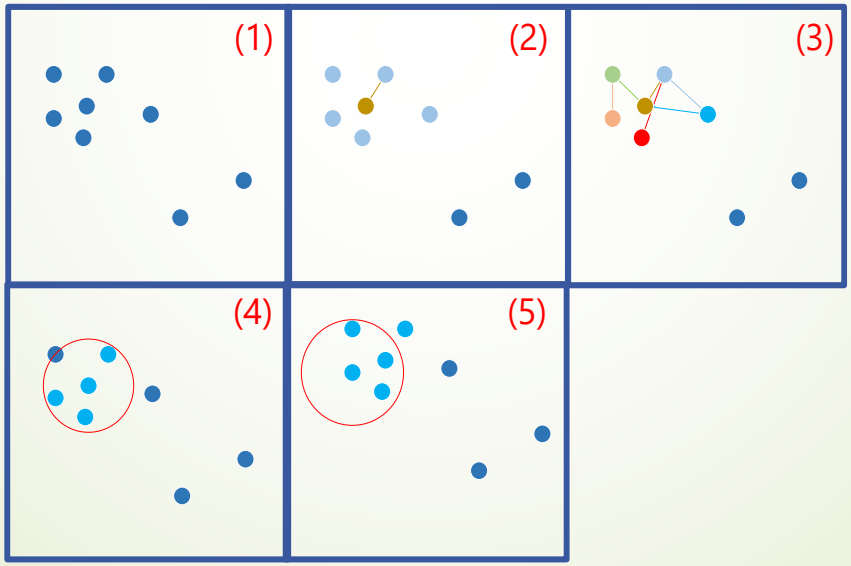
- تعیین شعاع گسترش با توجّه به چگالی محلّی

$$radius_{expansion} = max\left(Density(i), \frac{\sum_{j\ in\ knnbrs(i)} density(j)}{k}\right) \quad (4)$$



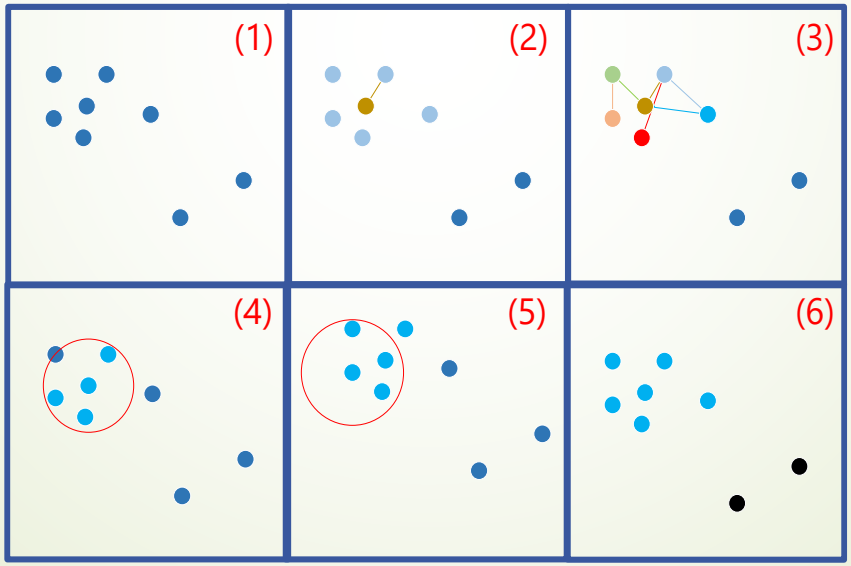
خوشه‌بندی محلّی KNNCA (ادامه)

- پویایی شعاع گسترش خوشه‌ها
- وابستگی شعاع گسترش به فاصله با همسایه‌ها



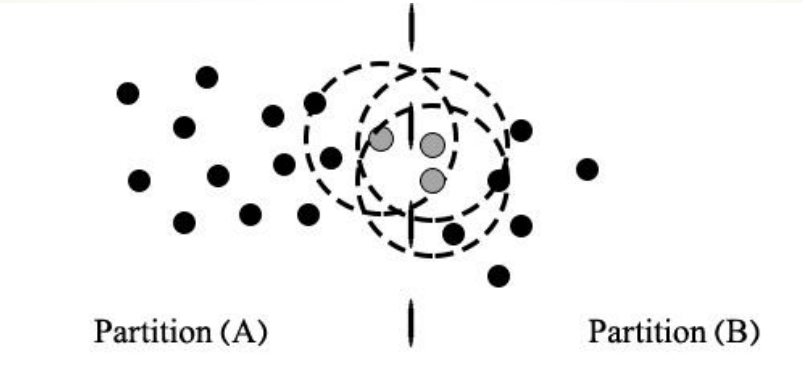
خوشه‌بندی محلّی KNNCA (ادامه)

- اختلاف چگالی زیاد نمونه‌های نویز با همسایه‌ها



ادغام نتایج

- تداخل شعاع گسترش نمونه‌های مرزی با مرز قسمت‌ها
- تجمیع نمونه‌های مرزی، شعاع گسترش و برچسب محلّی آن‌ها
- ادغام خوشهٔ هر نمونه با نمونه‌های داخل شعاع گسترش آن



مجموعه داده‌ها

- مجموعه داده‌های به کار رفته در ارزیابی خوشه‌بندی محلّی
  - داده‌های ساختگی

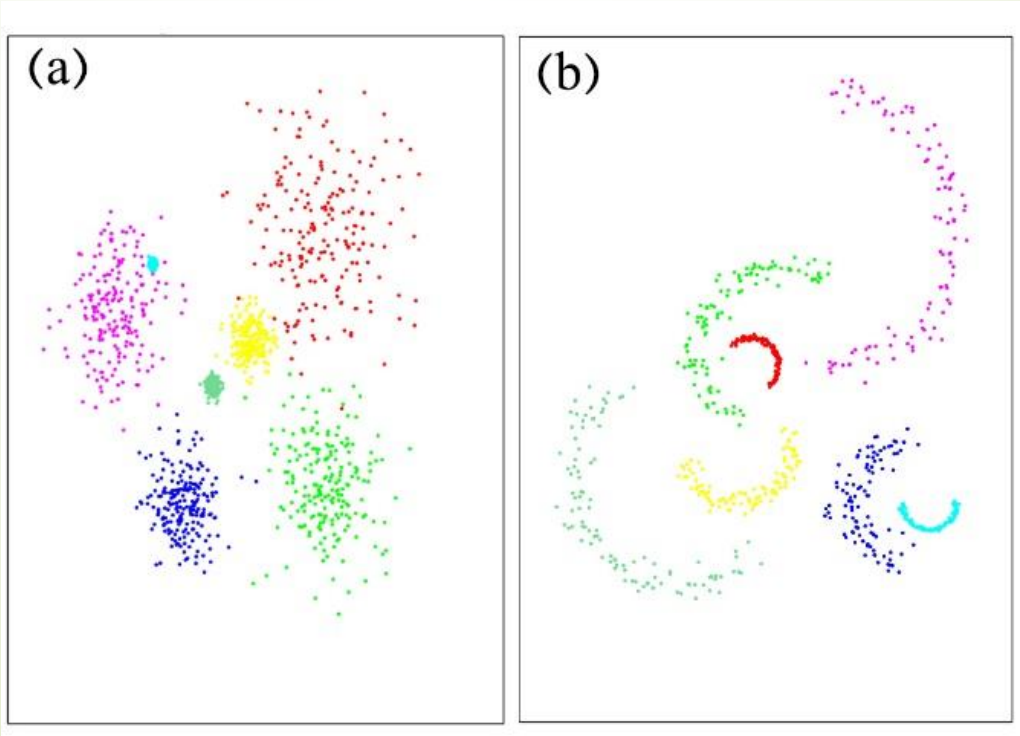
نام مجموعه داده	تعداد خصیصه‌ها	تعداد خوشه‌ها	تعداد نمونه‌ها	بیشترین اختلاف چگالی
آتش‌بازی	۲	۷	۱۴۰۰	۷۰ برابر
هلال‌ها	۲	۷	۷۰۰	۸ برابر

- مجموعه داده به کار رفته برای ارزیابی توزیع‌شده
- زیرمجموعه‌ای از kddcup99
- کاربرد تشخیص ناهنجاری

نام مجموعه داده	تعداد خصیصه‌ها	تعداد خوشه‌ها	تعداد نمونه‌ها
KDD99 Subset	۱۵	۲	۱۰۰۰۰

مجموعه داده‌های ساختگی

- تمرکز بر خوشه هایی با شکل دلخواه
- تمرکز بر خوشه‌هایی با چگالی متفاوت

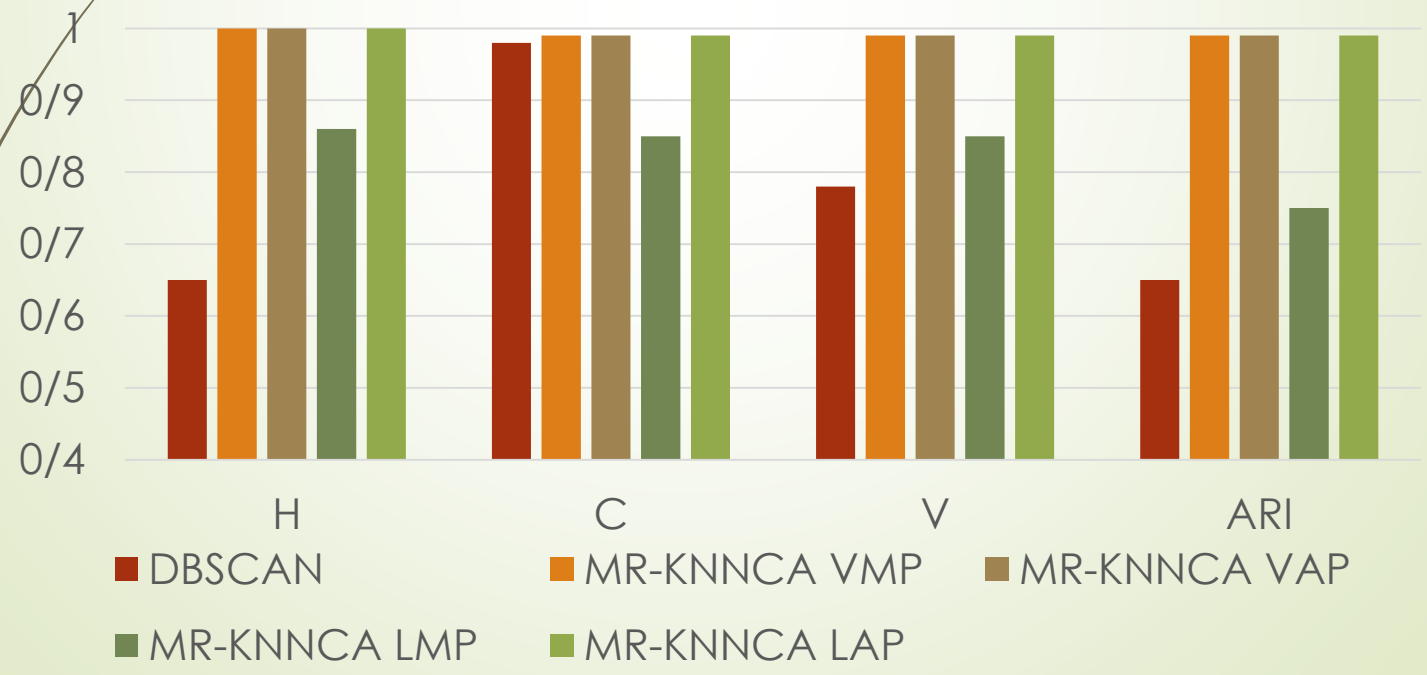


معیارهای ارزیابی

- شاخص تصادفی تعدیل‌شده (ARI)
  - نشان‌دهندهٔ میزان همگرایی نتایج در بازهٔ [۱, ۱-]
  - وابسته به برچسب خوشه‌بندی و برچسب واقعی **جفت نمونه‌ها**
- کامل بودن
  - نشانگر پخش شدن نمونه‌های یک کلاس در خوشه‌های مختلف در بازهٔ [۱ ,۰]
- همگن بودن
  - نشانگر خلوص خوشه‌ها در بازهٔ [۱ ,۰]
- امتیاز v
  - میانگین هارمونیک **همگن بودن** و **کامل بودن** نتایج خوشه‌بندی در بازهٔ [۱ ,۰]

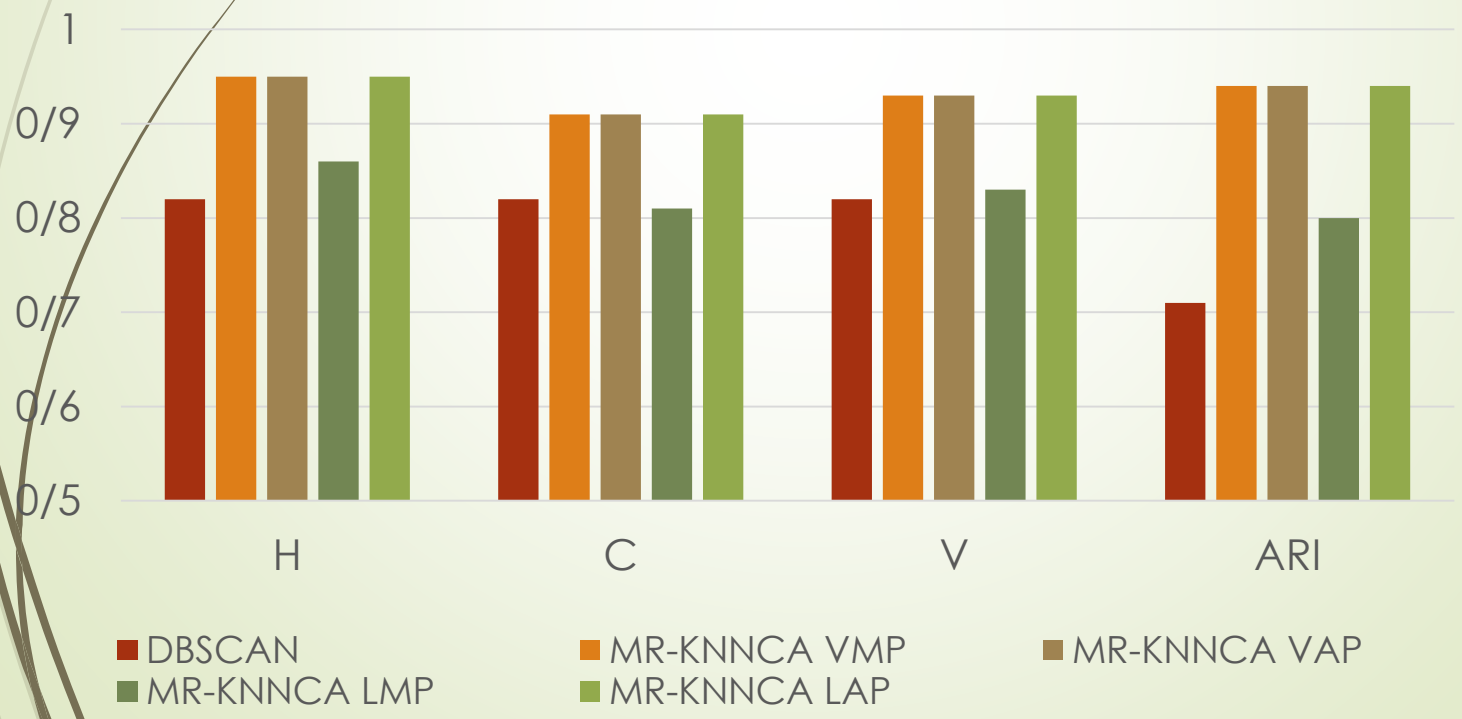
ارزیابی خوشه‌بندی داده‌های هلال‌ها

- امکان شناسایی خوشه‌هایی با چگالی متفاوت
- بهبود نتایج خوشه‌بندی با استفاده از آمارگان داده جهت تقسیم فضا



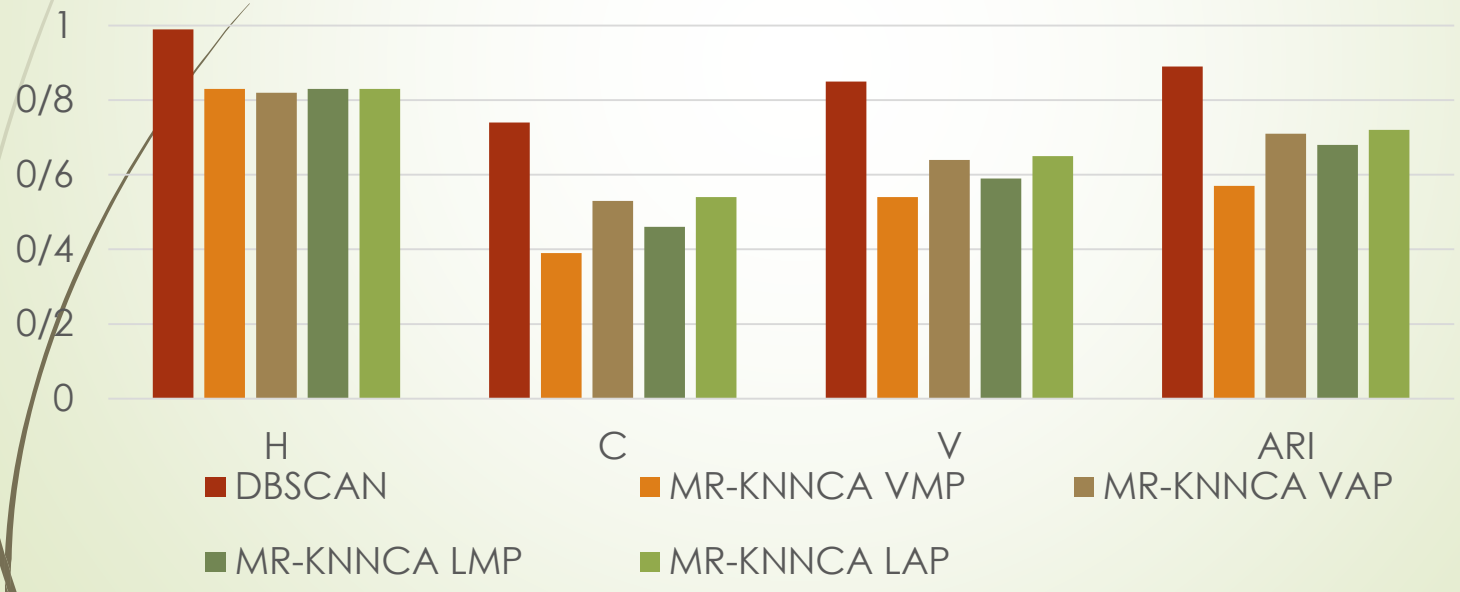
ارزیابی خوشه‌بندی داده‌های آتش‌بازی

- امکان شناسایی خوشه‌هایی با چگالی متفاوت
- بهبود نتایج خوشه‌بندی با استفاده از آمارگان داده جهت تقسیم فضا



ارزیابی خوشه‌بندی توزیع‌شده

- مقایسهٔ نتایج MR-KNNCA با DBSCAN برای داده‌های شبکه



مواردی که به آنها پرداختیم:

- ارائهٔ یک روش خوشه‌بندی KNNCA
- توسعهٔ روش خوشه‌بندی MR-KNNCA در چارچوب نگاشت‌کاهش
- استفاده از آمار موجود در داده‌ها جهت تقسیم فضا

کارهایی که می توان در آینده بیشتر به آنها پرداخت:

- استفاده از خاصیت انعطاف‌پذیری برای تغییر چارچوب خوشه‌بندی
- بهبود مرحلهٔ ادغام
- مقاوم‌سازی در برابر نویز
- استفاده از آمارگان بیشتر در تقسیم فضا

ارزیابی-معیار ارزیابی شاخص تصادفی تعدیل‌شده(ARI)

- نشانگر میزان شباهت بین دو انتساب برچسب (Hubert and Arabie 1985)
- a: تعداد جفت نمونه‌هایی با برچسب یکسان در هر دو انتساب
- b:تعداد جفت نمونه‌هایی با برچسب متفاوت در هر دو انتساب
- برچسب‌زنی تصادفی مقداری نزدیک به صفر خواهد داشت
- مقدار آن در بازهٔ [۱، ۱-] است

$$RI = \frac{a+b}{\frac{1}{2}(n-1)n}$$

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}$$

ارزیابی-معیار ارزیابی اطّلاعات متقابل تعدیل‌شده(AMI)

- نمایانگر میزان توافق بین دو مجموعه برچسب در بازهٔ [۰, ۱] (Vinh, Epps et al. 2010)
- مقداری نزدیک به صفر برای برچسب‌گذاری تصادفی، فارغ از تعداد خوشه‌ها و تعداد نمونه‌ها

$$H(U) = \sum_{i=1}^{|U|} P(i)\log(P(i))$$

$$MI(U,V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|} P(i,j)\log\left(\frac{P(i,j)}{P(i)\acute{P}(j)}\right)$$

- $P(i) = |U_i|/N$ بیانگر احتمال آن است که نمونه‌ای تصادفی از U به مجموعهٔ U_i متعلّق باشد
- $P(i,j) = |U_i \cap V_j|/N$ احتمال آن است که یک نمونهٔ تصادفی در هر دو مجموعهٔ i و j قرار بگیرد

$$AMI = \frac{MI - E[MI]}{\max\big(H(U), H(V)\big) - E[MI]}$$

ارزیابی-معیارهای همگن بودن، کامل بودن و امتیاز v

- معیارهای همگن بودن و کامل بودن نتایج خوشه‌بندی (Rosenberg and Hirschberg 2007)

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$H(C|K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{n_{c,k}}{n}.\log\left(\frac{n_{c,k}}{n}\right)$$

$$H(C) = -\sum_{c=1}^{|C|}\frac{n_c}{n}.\log\left(\frac{n_c}{n}\right)$$

- با فرض n تعداد کل نمونه‌ها و $n_{c,k}$ تعداد نمونه‌هایی از کلاس c که در خوشهٔ k قرار گرفته و n_c تعداد نمونه‌های متعلّق به کلاس c
- کامل بودن مشابه همگن بودن محاسبه می‌شود
- امتیاز v میانگین هارمونیک این دو است

$$v = 2 \times \frac{h \times c}{h + c}$$

ارزیابی-معیار ارزیابی ضریب سایه

- یک معیار داخلی ارزیابی که میزان تشابه نمونه‌های داخل خوشه و تفاوت خوشه‌ها با یکدیگر را می‌سنجد (Rousseeuw 1987)

$$s = \frac{b - a}{\max(a, b)}$$

- a میانگین فاصله با نمونه‌های خوشه
- b میانگین فاصله با نمونه‌های نزدیک‌ترین خوشه
- ضریب سایه برای کل نتایج، میانگین ضریب سایهٔ تک تک نمونه‌هاست
- مقدار ضریب سایه در بازهٔ [۱،۱-] قرار دارد