فرصت باقیمانده
۰۰ ۰۰ ۰۷ ۴۵ ۲۵
ثانیه دقیقه ساعت روز
۴۰٪ تخفیف جشنواره تخفیف بوک ماه مهر فرادرس کد تخفیف : MHR98
۵ مدم نمایش
منطقه
هشتگ‌ها
تخصص‌ها
جستجو

# روش های ارزیابی نتایج خوشه بندی — (Clustering Performance) معیارهای بیرونی (External Index)

۲۷ مرداد ۱۳۹۷ / ۷ دقیقه مطالعه / ایمان بلدیه

آموزش برنامه نویسی
آموزش متلب Matlab
نرم‌افزارهای مهندسی برق
نرم‌افزارهای مهندسی عمران
نرم‌افزارهای مهندسی مکانیک
نرم‌افزارهای مهندسی صنایع

دروس مهندسی برق
دروس مهندسی عمران
دروس مهندسی مکانیک
دروس مهندسی کامپیوتر
دروس مهندسی صنایع
دروس مهندسی معماری
دروس مهندسی الکترونیک
دروس مهندسی کنترل
دروس مهندسی قدرت

- آموزش‌های رایگان
- استخدام در فرادرس

آموزش اکسل Excel
آموزش نرم افزار SPSS
آموزش نرم افزار Catia
آموزش نرم افزار AutoCAD
آموزش نرم افزار ABAQUS

پربحث ترین مطالب

آخرین نظرات

- نوشین در فرمول نویسی در اکسل - آموزش مقدماتی نوشته است:

ممنون عالی بود.

- ایمان فروردین در چگونه یک اپلیکیشن هندسی با طراحی ترجمه کنیم؟ نوشته است:

سلام من با powdit شروع کردم به ترجمه افزونه oxygen ولی هیچ جایی که من ترجمه کردم فارسی نشده مشکل چه کنم وقت گذاشتم با افزونه loco translate هم شروع به ترجمه کردم دیدم اون رشته های ترجمه شده ...

- dorsa در تصاعد حسابی — به زبان ساده نوشته است:

mrc vaghean. sare class gg shode boodam vali alan ghashang ❤Mahmidam

کانال تلگرام «مجله فرادرس»

به کانال تلگرام ما بپیوندید تا از آموزش و مهارتی، مطالب منتشر شده در «بلاگ فرادرس» آگاه شوید.

عضویت

تطبیقی

هماهنگ‌ها

تفاضل‌ها

تعلق نقاط به هر خوشه، توسط عمل «برچسب‌گذاری» (Labeling) انجام می‌شود. با توجه به انواع مختلف، ویژگی‌ها برای نقاط، روش‌ها و الگوریتم‌های متنوعی نیز برای خوشه‌بندی وجود دارد. در نتیجه، ارزیابی نتایج خوشه‌بندی، ملاک است برای انتخاب الگوریتم مناسب. در این مطلب، شاخص‌های ارزیابی بیرونی معرفی و بررسی می‌شوند.

## روش‌های ارزیابی نتایج خوشه‌بندی با معیارهای بیرونی

از آنجایی که «تحلیل خوشه‌بندی» (Clustering Analysis) یک روش «بدون نظارت» (Unsupervised) محسوب می‌شود، روش‌های ارزیابی و سنجش کارایی آن با روش‌های «نظارت شده» (Supervised) تفاوت دارند.

- **فیلم آموزش خوشه‌بندی داده‌ها Clustering**
- **فیلم آموزش برنامه‌نویسی R - خوشه بندی سلسله مراتبی**
- **فیلم آموزش برنامه‌نویسی R - خوشه بندی تفکیکی**

معمولا ارزیابی نتایج خوشه‌بندی براساس نتایج فاصله و برچسب‌های تولید شده در تحلیل خوشه‌بندی انجام می‌شود. ارزیابی نتایج خوشه‌بندی می‌تواند از دو منظر بررسی شود: ارزیابی درونی (Internal Index) و ارزیابی بیرونی (External Index).

منظور از ارزیابی درونی، بررسی دستیابی به اهداف خوشه‌بندی است. همانطور که گفته شد هدف از خوشه‌بندی، ایجاد دسته‌هایی با بیشترین شباهت درونی و کمترین شباهت بین دسته‌ها است. در نتیجه در رویکرد ارزیابی درونی خوشه‌بندی از ترکیب نتوابع فاصله یا شباهت برای خوشه‌ها استفاده می‌شود. برای مثال اگر داده‌ها یک بعدی باشند، «تحلیل واریانس» (Analysis of Variance) می‌تواند نسبت میانگین تغییرات بین خوشه‌ها را به میانگین تغییرات درون خوشه‌ها اندازه‌گیری کرده و مشخص کند که آیا خوشه‌ها کاملا از یکدیگر جدا هستند یا خیر.

ولی در روش ارزیابی با معیارهای بیرونی، برای همه نقاط یک «برچسب واقعی» (Benchmark) وجود دارد که نشان دهنده تعلق نقاط به دسته‌ها است. معمولا به برچسب‌های واقعی، «استاندارد طلایی» (Gold Standard) نیز می‌گویند. از طرفی «برچسب‌های خوشه‌بندی» (Clustering Label) نیز که مربوط به خوشه‌ای است که یک نقطه درون آن قرار دارد. در این متن مجموعه دسته‌ها را با

$$S = \{S_1, S_2, \ldots, S_l\}$$

و خوشه‌ها را به صورت

$$C = \{C_1, C_2, \ldots, C_k\}$$

نشان می‌دهیم.

در روش ارزیابی بیرونی، مطابقت این دو گونه برچسب انجام می‌پذیرد. باید توجه داشت که ممکن است کدهای برچسب‌های حاصل از خوشه‌بندی با برچسب‌های واقعی یکسان نباشند. به این معنی که برچسب واقعی ۱ برای یک نقطه بیانگر متعلق بودن آن به دسته شماره ۱ است در حالیکه ممکن است شماره برچسب برای این نقطه در خوشه‌بندی برابر با ۴ باشد.

- **فیلم آموزش برنامه‌نویسی R - خوشه بندی سلسله مراتبی**
- **فیلم آموزش خوشه‌بندی داده‌ها Clustering**
- **فیلم آموزش برنامه‌نویسی R - خوشه بندی تفکیکی**

### مثال ۱

فرض کنید برای 5 نقطه برچسب واقعی و حاصل از خوشه‌بندی در جدول زیر قرار داشته باشد. هر چند برچسب‌ها یکسان نیستند ولی هر دو سری برچسب به دسته یا خوشه‌های یکسانی اشاره دارند.

| مقدار | 5 | 4 | 6 | 2 | 1 |
|---|---|---|---|---|---|
| برچسب واقعی | 1 | 1 | 1 | 2 | 2 |
| برچسب خوشه | 2 | 2 | 2 | 1 | 1 |

پس هنگام مقایسه برچسب‌ها، باید توجه داشت که عدم مطابقت برچسب‌های واقعی با خوشه‌های ممکن است حاصل از این مسئله باشد. در ادامه به بررسی روش‌هایی می‌پردازیم که با در نظر گرفتن این مسئله، مطابقت برچسب‌های حاصل از خوشه‌بندی را با واقعی ارزیابی می‌کنند.

## شاخص خلوص

یکی از ساده‌ترین شاخص‌های ارزیابی بیرونی در خوشه‌بندی، «شاخص خلوص» (Purity Index) است که درصد مطابقت بین برچسب‌های خوشه‌بندی و برچسب‌های واقعی را می‌سنجد. در این حالت برچسب هر خوشه با برچسب واقعی دسته‌ای که بیشترین اشتراک را دارد، مطابقت پیدا کرده و تعداد نقاطی از خوشه که در دسته صحیح طبقه‌بندی شده‌اند شمارش می‌شود. نسبت این تعداد به تعداد کل نقاط شاخص خلوص را می‌سازد و شکل محاسباتی آن به صورت زیر است:

$$Purity(S, C) = \frac{\sum_{i=1}^{k} \max_j |S_m \cap C_k|}{N}$$

باید توجه داشت که منظور از

$$|S_m \cap C_k|$$

تعداد نقاط مشترک از خوشه

$$C_k$$

با دسته

$$S_m$$

و N نیز تعداد کل نقاط است. با توجه به شیوه محاسبه شاخص خلوص، مشخص است که حداکثر مقدار برای آن، ۱ خواهد بود و این در زمانی اتفاق می‌افتد که برچسب‌های حاصل از خوشه‌بندی، کاملا با برچسب‌های واقعی مطابقت داشته باشند. همچنین اگر هیچ برچسب خوشه‌ای با برچسب واقعی مطابقت نداشته باشد، این شاخص صفر می‌شود.

### مثال 2

فرض کنید برای 5 نقطه، برچسب‌های واقعی و برچسب‌های حاصل از تحلیل خوشه‌بندی در جدول زیر قرار دارند.

| خوشه | دسته | مقدارها |
|---|---|---|
| ۲ | ۱ | ۵ |
| ۲ | ۱ | ۴ |
| ۲ | ۲ | ۶ |
| ۱ | ۲ | ۲ |
| ۱ | ۳ | ۱ |

به منظور محاسبه شاخص خلوص، جدول متقاطعی از برچسب‌ها ایجاد می‌کنیم و محاسبات را طبق فرمول انجام می‌دهیم. شاخص شفافیت برای خوشه‌های ایجاد شده برابر است با 0.6 یا ۶۰٪:

| | دسته ۱ | دسته ۲ | دسته ۳ | مقدار حداکثر |
|---|---|---|---|---|
| خوشه ۱ | ۰ | ۱ | ۱ | ۱ |
| خوشه ۲ | ۲ | ۱ | ۰ | ۲ |
| | | | جمع | ۳ |
| شاخص شفافیت | | | نسبت | ۰.۶ |

**نکته:** در این روش احتمالی نیست که تعداد دسته‌ها با خوشه‌ها برابر باشد. به ماتریس بالا که بنگرید وضعیتی را نشان می‌دهد دقت کنید.

### خصوصیات شاخص خلوص

**سادگی در محاسبات:** با توجه به شیوه محاسبه این شاخص، محاسبات طولانی برای بدست آوردن میزان مطابقت لازم نیست.

**مستقل از تعداد خوشه‌ها:** شاخص خلوص به تعداد خوشه‌ها توجه ندارد. در نتیجه نمی‌توان این شاخص را به عنوان معیاری برای سنجش مطابقت تعداد خوشه‌ها نیز در نظر گرفت.

**کاهش کارایی با افزایش تعداد خوشه‌ها:** اگر تعداد خوشه‌ها زیاد باشد و هیچ هماهنگی نیز بین برچسب‌های واقعی و خوشه‌های وجود نداشته باشد ممکن است شاخص خلوص به ۱ نزدیک شود که یک عیب برای چنین شاخص است.

## شاخص زند اصلاح شده

برای نشان دادن میزان شباهت بین دو شیوه برچسب‌گذاری می‌توان از «شاخص زند» (Rand Index) استفاده کرد. این شاخص توسط دانشمند آمار ویلیام زند (William Rand) در سال 1971 در مقاله‌ای با عنوان «معیارهای هدف برای ارزیابی روش‌های خوشه‌بندی» (Objective criteria for the evaluation of clustering methods) معرفی شد.

برای محاسبه آن باید دو پارامتر را اندازه‌گیری کنیم:

- **فیلم آموزش خوشه‌بندی داده‌ها Clustering**
- **فیلم آموزش برنامه‌نویسی R - خوشه بندی سلسله مراتبی**
- **فیلم آموزش برنامه‌نویسی R - خوشه بندی تفکیکی**

A: تعداد زوج‌هایی که هم در خوشه‌ها و هم در دسته‌ها در کنار هم هستند. به بیان دیگر هم در خوشه‌ها دارای برچسب یکسانی هستند و هم برچسب دسته‌ها برای آن‌ها یکسان است.

B: تعداد زوج‌هایی که هم در خوشه‌ها و هم در دسته‌ها از یکدیگر جدا هستند. یعنی برچسب خوشه‌هایشان متفاوت است و البته برچسب دسته‌های متفاوتی نیز دارند.

مثلا در جدول اطلاعاتی مربوط به مثال ۳، زوج‌های

$$(4, 5)$$

در یک دست و در یک خوشه هستند. هرچند ممکن است شماره برچسب‌های متفاوتی برای خوشه‌ها یا دسته‌ها وجود داشته باشد. همچنین زوج

$$(5, 2)$$

نیز در دو دسته جدا قرار داشته و ضمناً برچسب خوشه‌های جداگانه نیز دارند.

حال برای محاسبه شاخص زند کافی است که حاصل جمع A و B را، با تعداد کل زوج‌ها تقسیم کنیم.

$$Rand(S, C) = \frac{A + B}{\binom{n}{2}} = \frac{A + B}{N(N-1)/2}$$

### مثال ۳

اگر برای داده‌های مربوط به مثال ۳، شاخص Rand را محاسبه کنیم، مطابقت برچسب‌های واقعی و خوشه‌ها را به صورت زیر می‌توان مشاهده کرد.

$$Rand(S, C) = 0.6$$

| | |
|---|---|
| A | ۱ |
| B | ۵ |
| جمع | ۶ |
| Rand | ۰.۶ |

اگر خوشه‌ها مطابق با دسته‌ها ایجاد شده باشند، شاخص زند برابر با ۱ خواهد بود. ولی اگر خوشه‌بندی به صورت تصادفی ایجاد شده باشد، دلیلی ندارد که مقدار این شاخص برابر با صفر باشد. برای رفع این مشکل از «شاخص زند اصلاح شده» (Adjusted Rand Index) استفاده می‌شود. این شاخص را به صورت ARI نشان داده و به شکل زیر محاسبه می‌شود:

$$ARI(S, C) = \frac{Rand(S, C) - E(Rand(S, C))}{max(Rand(S, C)) - E(Rand(S, C))}$$

منظور از E نیز امید-ریاضی شاخص زند است.

### خصوصیات شاخص زند اصلاح شده

حدود مقدارها: ARI مقداری بین ۱ و ۱- خواهد بود. در حالتی که ARI=1 باشد، مطابقت کامل بین برچسب‌های واقعی و خوشه‌های وجود دارد و در مقابل اگر مقدار این شاخص برابر با ۱- باشد نشانگر برچسب‌گذاری تصادفی در حین خوشه‌بندی است.

- **شاخص کارا برای مقایسه چندین روش:** از آنجایی که این شاخص به توافق بین برچسب‌ها

۴۰٪ تخفیف جشنواره تخفیف بوک ماه مهر فرادرس کد تخفیف: MHR98
فرصت باقیمانده
روز ساعت دقیقه ثانیه
دوره‌ها
هاکاتون‌ها
مقاله‌ها
تماس با ما
دوره‌ها
هاکاتون‌ها
مقاله‌ها
تماس با ما
© فرادرس ۱۳۹۷-۱۳۹۹. تمامی حقوق محفوظ است.