مجله‌ی بررسی‌های آمار رسمی ایران
سال۲۴، شماره‌ی ۲، پاییز و زمستان ۱۳۹۲، صص ۲۲۷– ۲۴۰

خوشه‌بندی سلسله‌مراتبی و K-میانگین در نرم‌افزارهای R، SAS و MATLAB

حسین هشیارمنش[†,*]، میلاد فرهادی[‡]، علی هشیارمنش[★] و نگار جعفریان[¥]

[†] پژوهشکده‌ی آمار
[‡] مؤسسه‌ی آموزش عالی صدرا
[★] دانشکده‌ی کشاورزی و منابع طبیعی دانشگاه آزاد اسلامی واحد کرج
[¥] دانشگاه صنعتی امیرکبیر

چکیده: خوشه‌بندی سلسله‌مراتبی و خوشه‌بندی K-میانگین، با توجه به توانایی‌هایی که در برطرف‌کردن نیازها و مشکلات پژوهشی در علوم مختلف دارند، کاربرد وسیع و گسترده‌ای در بین پژوهشگران پیدا کرده‌اند. کاربردهای وسیع خوشه‌بندی در زمینه‌های مختلف بهداشت و درمان، فنی و مهندسی، علوم اجتماعی و انسانی سبب رشد علم در این زمینه‌ها شده است. نرم‌افزارهای R، SAS و MATLAB به‌دلیل قابلیت‌های برنامه‌نویسی‌ای که دارند، از پرکاربردترین نرم‌افزارها در تجزیه و تحلیل خوشه‌بندی هستند. ولی با این وجود هر کدام از این نرم‌افزارها قابلیت و محدودیت‌های خاصی برای خوشه‌بندی دارند. به‌همین دلیل آشنایی پژوهشگران با نحوه‌ی خوشه‌بندی در هر یک از این نرم‌افزارها این امکان را به آن‌ها می‌دهد تا با توجه به نوع داده‌هایی که در اختیار دارند و نیازهایشان از نرم‌افزاری که راحت‌تر و سریع‌تر نیازهای آن‌ها را برطرف می‌کند، برای اجرای خوشه‌بندی سلسله‌مراتبی و K-میانگین استفاده کنند.

واژگان کلیدی: خوشه‌بندی سلسله‌مراتبی؛ خوشه‌بندی K-میانگین؛ نرم‌افزارهای R، SAS و MATLAB.

---

* نویسنده‌ی عهده‌دار مکاتبات
دریافت: ۱۳۹۳/۲/۳۱، پذیرش: ۱۳۹۳/۹/۳.

۱- مقدمه

با توجه به گسترش علوم مختلف، برخی از تکنیک‌های یک علم به‌دلیل توانایی در برطرف کردن نیازهای علوم مختلف، کاربرد وسیع و گسترده‌ای، پیدا کرده‌اند. یکی از این تکنیک‌ها خوشه‌بندی[1] است. خوشه‌بندی به‌دلیل توانایی و قابلیت‌های بالایی که در تلخیص اطلاعات و دسته‌بندی آن‌ها دارد مورد توجه محققان و پژوهشگران علوم مختلف قرار گرفته است.

در مهندسی از خوشه‌بندی برای فشرده‌سازی تصویر، فشرده‌سازی صوت، تشخیص گوینده‌ی گفتار، تشخیص چهره‌ی افراد، بازیابی عکس، تحلیل سیگنال رادار، شناسایی پارازیت، تقسیم‌بندی و تحلیل تصاویر ماهواره‌ای و تصاویر پزشکی استفاده می‌کنند. در پزشکی از خوشه‌بندی برای شناسایی پروتئین‌ها، ژن‌ها، عوامل بیماری‌زا، میکروب‌ها و برای شناسایی راه‌های درمان بیماری‌ها و تهیه‌ی دارو استفاده می‌کنند. در اقتصاد از خوشه‌بندی برای شناسایی الگوی خرید، چگونگی سرمایه‌گذاری، دسته‌بندی عوامل ثروت و فقر، استفاده می‌کنند [۳] و [۵].

به‌دلیل فراگیر شدن استفاده از تکنیک‌های خوشه‌بندی در علوم مختلف، نیاز به یادگیری خوشه‌بندی برای پژوهشگران وجود دارد. پژوهشگران اغلب به یک یا در موارد محدودی به دو نرم‌افزار تسلط دارند. ما می‌خواهیم دستورهای خوشه‌بندی برای دو مدل خوشه‌بندی سلسله‌مراتبی[2] [۱] و خوشه‌بندی K-میانگین[3] [۱۲] که کاربرد وسیعی در علوم مختلف دارند [۶]، [۷]، [۹] و [۱۳]، را در نرم‌افزارهای R، SAS و MATLAB، [۲]، [۸]، [۱۰] و [۱۱] که از پر کاربردترین نرم‌افزارها در تجزیه و تحلیل آماری هستند، معرفی کنیم.

برای برنامه‌نویسی می‌توان مراحل اساسی خوشه‌بندی را شامل قسمت‌های

۱. آماده‌سازی داده‌ها برای خوشه‌بندی
۲. دستورهای خوشه‌بندی
۳. رسم نمودار خوشه‌بندی مشاهدات

در نظر گرفت. در قسمت‌های بعد کدهای برنامه‌نویسی برای هر یک از مراحل اساسی در خوشه‌بندی سلسله‌مراتبی و خوشه‌بندی K-میانگین، در نرم‌افزارهای R، SAS و MATLAB معرفی شده است.

۲- خوشه‌بندی سلسله‌مراتبی

خوشه‌بندی سلسله‌مراتبی یکی از پرکاربردترین روش‌های خوشه‌بندی است. در این روش ابتدا فاصله دو به دوی مشاهدات از هم محاسبه می‌شود. پس از تعیین فاصله دو به دوی مشاهدات، با توجه به نزدیکی مشاهدات نسبت به یکدیگر، مشاهدات با هم تشکیل یک خوشه‌ی جدید می‌دهند. این کار تا جایی پیش می‌رود که تمام مشاهدات در یک خوشه قرار می‌گیرند.

خوشه‌بندی سلسله‌مراتبی چگونگی ترتیب مشاهدات و خوشه‌ها با هم را، به‌صورت نمودار درختی نمایش می‌دهد. می‌توان به کمک نمودار درختی، نحوه و ترتیب خوشه‌ها را مشخص کرد.

از مزیت‌های خوشه‌بندی سلسله‌مراتبی، سادگی و قابلیت درک برای تمامی پژوهشگران است. این روش شامل مدل‌های متنوعی است که می‌تواند نیازهای متعددی را رفع کند.

داده‌هایی که بر روی آن تکنیک‌های خوشه‌بندی را اجرا می‌کنیم، داده‌های جمع‌اوری‌شده به روش ثبتی، و یا داده‌هایی شبیه‌سازی‌شده هستند. در آزمایشگاه‌های آمار برای آموزش و یا بررسی کارایی و دقت یک روش و یا مدل از روش‌های شبیه‌سازی داده‌ها استفاده می‌کنیم. در ادامه ما از دو مجموعه داده، که هم از داده‌های جمع‌اوری‌شده به روش ثبتی و هم از داده‌های شبیه‌سازی‌شده است، استفاده می‌کنیم. داده‌های جمع‌اوری‌شده به روش ثبتی مورد استفاده در این مقاله، داده‌هایی با عنوان فقر است. این مجموعه داده شامل چهار متغیر نام کشور، تعداد مرگ و میر نوزادان، تعداد مرگ و میر، و تعداد تولد است. مجموعه داده‌های شبیه‌سازی‌شده، شامل ۱۰۰ مشاهده از یک متغیر آمیخته است. متغیر آمیخته‌ای که مجموعه داده‌های آن با احتمال ۰٫۶ از دو توزیع $N(۵, ۱)$ و $N(۱۰, ۱)$ است. در اینجا تولید داده از توزیع آمیخته، برای بررسی توانایی تکنیک خوشه‌بندی در جداسازی داده‌های آمیخته است. در مرحله‌ی اول از مراحل اساسی خوشه‌بندی، ابتدا مجموعه داده‌های شبیه‌سازی‌شده را وارد نرم‌افزارها می‌کنیم. در هریک از نرم‌افزارهای R، SAS و MATLAB، ابتدا مجموعه داده‌ای ۱۰۰تایی با نام label، از توزیع دوجمله‌ای با احتمال ۰٫۶ تولید می‌شود. بر اساس مشاهدات مجموعه داده‌ی label، مجموعه داده‌ی آمیخته‌ی X، تولید می‌شود. اگر مشاهده‌ای در مجموعه داده‌ی label برابر

با ۱ باشد، عددی از توزیع $N(۵, ۱)$ و اگر مشاهده‌ای در مجموعه داده‌ی label برابر با صفر باشد، عددی از توزیع $N(۱۰, ۱)$، برای مجموعه داده‌ی آمیخته‌ی X تولید می‌شود. دستورهایی که برای این منظور در هر یک از نرم‌افزارهای R، SAS و MATLAB مورد استفاده قرار می‌گیرند، در شکل ۱ آمده است. بخشی از داده‌های تولید شده از دستورهای موجود در شکل ۱، در پیوست آمده است.

مجموعه داده‌ای که می‌خواهیم بر روی آن‌ها خوشه‌بندی سلسله‌مراتبی انجام دهیم می‌تواند به‌صورت تصادفی تولید شود و یا از داده‌های جمع‌اوری‌شده به روش ثبتی باشد. دستورهایی که برای ورود داده‌های جمع‌اوری‌شده به روش ثبتی در هر یک از نرم‌افزارهای R، SAS و MATLAB مورد استفاده قرار می‌گیرند در شکل ۲ آمده است.

در مرحله‌ی دوم از مراحل اساسی خوشه‌بندی، برای خوشه‌بندی سلسله‌مراتبی مشاهدات، دو انتخاب مهم باید انجام داد. این دو انتخاب عبارت‌اند از: تعیین متری برای محاسبه‌ی فاصله دو به دوی بین مشاهدات و دیگری تعیین روش محاسبه‌ی فاصله‌ی خوشه‌های ادغام‌شده با هم.

R

```
label=rbinom(100,1,0.6)
x=rnorm(100,ifelse(label>0,5,10),1)
```

SAS

```
data section;
    do i=1 to 100;
    label=ranbin(10,1,0.6);
        if label>0 then x= rand("Normal", 5, 1);
        if label=0 then x= rand("Normal", 10, 1);
    output; end;
run;
```

MATLAB

```
lable= binornd (1,0.6,[100,1]);
    for i=1:n
        if lable(i)>0
            x(i)=normrnd(5,1);
        else x(i)= normrnd(10,1);
        end
    end
```

شکل ۱- کدهای برنامه‌نویسی تولید عدد تصادفی برای خوشه‌بندی

R

```
x<-read.table("c:\\Real-Data-Poverty.txt",header=T)
j.mat<-as.matrix(x[1:3])
```

SAS

```
PROC IMPORT OUT= WORK.x
            DATAFILE= "C:\Real-Data-Poverty.txt"
            DBMS=TAB REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;
```

MATLAB

```
x = importdata('C:\Real-Data-Poverty.txt');
```

شکل ۲. دستورهای ورود داده‌های جمع‌اوری‌شده به روش ثبتی به نرم‌افزار برای خوشه‌بندی

به‌طور شهودی برای مشاهدات $X^1, \ldots, X^n$، که $X^k = X_1^k, \ldots, X_m^k$، m متغیر اندازه‌گیری‌شده، ثبت گردیده است. برای خوشه‌بندی این مشاهدات با توجه به متغیرهای اندازه گیری‌شده برای هر مشاهده، فاصله‌ی بین مشاهدات را با متری که معمولاً متر اقلیدسی[۴] [۴] است اندازه می‌گیرند. متر اقلیدسی $d(x^i, x^j)$ فاصله‌ی بین دو مشاهده‌ی x^i و x^j را محاسبه می‌کند. روش محاسبه‌ی فاصله‌ی دو مشاهده با استفاده از متر اقلیدسی به‌صورت زیر می‌باشد [۱]:

$$d(x^i, x^j) = \sqrt{\left(x_1^i - x_1^j\right)^2 + \cdots + \left(x_m^i - x_m^j\right)^2}$$

با استفاده از این روش فاصله بین مشاهدات را به‌صورت دو به دو محاسبه می‌کنیم و فاصله‌های به‌دست آمده را در ماتریس فاصله قرار می‌دهیم. بعد از تشکیل ماتریس فاصله، کوچک‌ترین عدد در ماتریس فاصله را پیدا می‌کنیم (به غیر از عناصر روی قطر اصلی)

عدد فوق در درایه‌ای است که محل تقاطع دو مشاهده می‌باشد. این عدد نشان‌دهنده‌ی این است که دو مشاهده‌ی فوق کم‌ترین فاصله‌ی بین مشاهدات را از یکدیگر دارند، لذا می‌توان دو مشاهده را در قالب یک خوشه قرار داد.

در این‌جا این سؤال مطرح می‌شود که وقتی دو مشاهده با هم تشکیل یک خوشه‌ی جدید می‌دهند، چگونه می‌توان فاصله‌ی این خوشه با دیگر مشاهدات را محاسبه کرد؟

فرض کنید در ماتریس فاصله، کم‌ترین عدد مخصوص به درایه‌ای باشد که محل تقاطع دو مشاهده‌ی a و b است. حال سؤال این است، هنگامی که دو مشاهده‌ی a و b تشکیل خوشه‌ی جدید ab را می‌دهند، فاصله‌ی خوشه‌ی جدید تا مشاهدات دیگر (به‌عنوان مثال مشاهده‌ی d)، چگونه محاسبه می‌شود؟

کدهای برنامه‌نویسی که برای این منظور در هر یک از نرم‌افزارهای R، SAS و MATLAB مورد استفاده قرار می‌گیرند، در شکل ۳ آمده است.

در این قسمت برای جلوگیری از بروز خطا در تجزیه و تحلیل باید به نوع داده‌ها و تحلیلی که قرار است انجام شود بسیار دقت کرد. از پرکاربردترین روش‌ها برای تعیین فاصله‌ی خوشه‌ی جدید ab تا مشاهدات دیگر، در خوشه‌بندی سلسله‌مراتبی: روش نزدیک‌ترین فاصله[۵]، روش دورترین فاصله[۶]، روش پیوند میانگین[۷]، روش پیوند میانه[۸] و روش پیوند به مرکز[۹] است. در هر یک از نرم‌افزارهای R، SAS و MATLAB تعیین روش مورد نظر، در بخش method، انجام می‌شود (شکل ۳). انتخاب هر یک از این روش‌ها بر مبنای نوع داده، و تحلیلی که قرار است روی آن صورت گیرد تعیین می‌شود. برای بررسی این روش‌ها و درک بهتر از تفاوت هر یک از آن‌ها می‌توان به [۱] مراجعه کرد.

مرحله‌ی سوم از مراحل اساسی خوشه‌بندی، رسم نمودار خوشه‌بندی است. دستورهایی که برای رسم نمودار خوشه‌بندی سلسله‌مراتبی در هر یک از نرم‌افزارهای R، SAS و MATLAB مورد استفاده قرار می‌گیرند در شکل ۴ آمده است.

خوشه‌بندی سلسله‌مراتبی در داده‌هایی با بُعد بالا، به‌دلیل زمان زیادی که صرف محاسبات می‌نماید، مقرون به صرفه و گاهی نیز قابل محاسبه نیست. البته این مشکل را با خوشه‌بندی K-میانگین رفع می‌کنند.

در قسمت بعد به معرفی خوشه‌بندی K-میانگین، و دستورهای برنامه‌نویسی آن در نرم‌افزارهای R، SAS و MATLAB می‌پردازیم.

R

```
j.mat<-as.matrix(x)
x.dist<-dist(j.mat)
h<-hclust(x.dist,method="ave")
```

SAS

```
proc cluster data=section method=ward
          ccc pseudo print=15 outtree=Tree;
      var x;
      id i;
run;
```

MATLAB

```
e = linkage(x,'ward','euclidean');
```

شکل ۳. کدهای برنامه‌نویسی خوشه‌بندی سلسله‌مراتبی داده‌ها

R

```
plot(h, hang =-1,col="blue")
```

SAS

```
proc tree data=Tree out=New nclusters=2
        graphics haxis=axis1 horizontal;
      height _rsq_;
      copy x ;
      id i;
run;
```

MATLAB

```
dendrogram(e)
```

شکل ۴. کدهای برنامه‌نویسی رسم نمودار خوشه‌بندی سلسله مراتبی داده‌ها

۳- خوشه‌بندی K-میانگین

در روش خوشه‌بندی سلسله‌مراتبی با زیاد شدن تعداد مشاهدات، تعداد محاسبات بیش‌تر می‌شود. این امر سبب زمان‌گیرشدن و در بعضی مواقع عدم دست‌یابی به نتیجه می‌شود. برای رفع این مشکل از خوشه‌بندی K–میانگین استفاده می‌شود.

در خوشه‌بندی K–میانگین، ابتدا نقاطی به‌عنوان مرکزیت تعیین می‌شود. تعداد این نقاط با توجه به تعداد خوشه‌هایی که وجود دارد، تعیین شده است. پس از تعیین نقاطی به‌عنوان مرکز، فاصله‌ی هر نقطه تا مراکز، تعیین می‌شود، سپس نزدیک‌ترین نقاط به هر مرکز، با هم تشکیل خوشه می‌دهند.

در ادامه به معرفی کدهای برنامه‌نویسی در مراحل اساسی خوشه‌بندی K–میانگین، می‌پردازیم.

کدهای برنامه‌نویسی، برای مرحله‌ی اول از مراحل اساسی در خوشه‌بندی K–میانگین، یعنی نحوه‌ی ورود داده‌های شبیه‌سازی‌شده و یا فایل داده‌های جمع‌اوری‌شده به روش ثبتی در نرم‌افزارهای R، SAS و MATLAB مانند دستورهای آمده در شکل‌های ۱ و ۲ است.

در مرحله‌ی دوم از مراحل اساسی خوشه‌بندی K–میانگین، دو انتخاب مهم باید انجام داد. اولین انتخاب، تعیین تعداد خوشه‌ها است. دومین انتخاب، تعیین حداکثر تعداد تکرارها برای رسیدن به خوشه‌هایی همگن است.

در اولین انتخاب، K نقطه، به‌عنوان مراکز هر خوشه تعیین می‌شود. سپس فاصله‌ی مشاهدات تا هر یک از مراکزی که مشخص شده، محاسبه می‌شود. مشاهداتی که کم‌ترین فاصله را تا هر یک از مراکز دارند، با هم تشکیل خوشه می‌دهند.

در خوشه‌بندی K–میانگین بر خلاف شیوه‌های سلسله‌مراتبی امکان جابه‌جایی مشاهده‌ای از یک خوشه به خوشه دیگر وجود دارد. در دومین انتخاب، میانگین خوشه‌هایی که در مرحله‌ی قبل محاسبه شده است را به‌عنوان مرکز در نظر گرفته، دوباره فاصله‌ی مشاهدات تا هر یک از مراکز جدید را محاسبه کرده، خوشه‌های جدید را تشکیل می‌دهیم. تعداد این تکرارها را در دومین انتخاب تعیین می‌کنیم. یکی از دلایل این امر در خوشه‌بندی K–میانگین، کم کردن فاصله‌ی مشاهدات از مرکز خوشه، برای رسیدن به کم‌ترین واریانس درون خوشه‌ای است. دستورهایی که برای این منظور در هر یک از نرم‌افزارهای R، SAS و MATLAB مورد استفاده قرار می‌گیرند، در شکل ۵ آمده

است. در این برنامه تعداد خوشه‌ها برابر با ۲ و تعداد تکرار (دومین انتخاب) برابر با ۸۵ تعیین شده است.

R

```
km <- kmeans(x, 2, iter.max =85)
```

SAS

```
proc fastclus data=section outseed=pish1 out=out
    maxclusters=2 maxiter=85 summary;
    var x;
run;
```

MATLAB

```
opts = statset('MaxIter', 85, 'Display', 'iter');
[G,C] = kmeans(X, K, 'options',opts,'distance'
               ,'sqEuclidean', 'start','sample');
```

شکل ۵- کدهای برنامه‌نویسی خوشه‌بندی K–میانگین داده‌ها

R

```
plot(x, col = km$cluster)
points(km$centers, col = 1:2, pch = 8)
```

SAS

```
proc sgplot;
    scatter y=x x=x / group=cluster;
run;
```

MATLAB

```
plot(X(G==1,1),'r.','MarkerSize',12)
plot(X(G==2,1),'b.','MarkerSize',12)
plot(C(:,1),'kx','MarkerSize',12,'LineWidth',2)
plot(C(:,1),'ko','MarkerSize',12,'LineWidth',2)
legend('Cluster 1','Cluster 2','Centroids','Location','NW')
```

شکل ۶- کدهای برنامه‌نویسی رسم نمودار خوشه‌بندی K–میانگین داده‌ها

۴- بحث و نتیجه‌گیری

خوشه‌بندی یکی از روش‌های تلخیص مشاهدات است. خوشه‌بندی در مواردی مورد استفاده قرارمی‌گیرد که پژوهشگر، ایده‌ای برای طبقه‌بندی و دسته‌بندی مشاهدات در اختیار ندارد. در بین مدل‌های خوشه‌بندی، خوشه‌بندی سلسله‌مراتبی به‌عنوان تکنیکی که دقت بالایی دارد و خوشه‌بندی K-میانگین به‌عنوان یکی از ساده‌ترین تکنیک‌های خوشه‌بندی به‌طور گسترده مورد استفاده‌ی پژوهشگران در رشته‌های مختلف قرار می‌گیرند.

از آن‌جایی که نرم‌افزارهای تحلیلی دارای برنامه‌ای از پیش نوشته شده هستند، برای خوشه‌بندی مشاهدات در موارد کم‌تری مورد استفاده قرار می‌گیرند. لذا پژوهشگران برای جبران محدودیت‌های نرم‌افزارهای تحلیلی، از نرم‌افزارهایی که قابلیت برنامه‌نویسی دارند استفاده می‌کنند. مطرح‌ترین این نرم‌افزارها R، SAS و MATLAB هستند. ولی هر یک از این نرم‌افزارها نیز به تنهایی دارای نقاط قوت و ضعف هستند. پژوهشگران زیادی تنها به یکی از این نرم‌افزارها تسلط دارند. این امر سبب می‌شود چنان‌که پژوهشگری بنا به دلایلی خاص، از جمله هزینه‌بر بودن استفاده از نرم‌افزار و یا توانایی نرم‌افزار به انجام محاسبات سنگین، بنا باشد با نرم‌افزار دیگری، غیر از نرم‌افزاری که به آن تسلط دارد برنامه‌ی خوشه‌بندی را بنویسد، محتاج آموزش آن نرم‌افزار می‌باشد. ما سعی کرده‌ایم دستورهای لازم برای خوشه‌بندی سلسله‌مراتبی وخوشه‌بندی K-میانگین را با نرم‌افزارهای برنامه‌نویسی پرکاربرد معرفی کنیم.

در هر خوشه‌بندی می‌توان مراحل آماده‌سازی داده‌ها برای خوشه‌بندی، دستورهای خوشه‌بندی، و رسم نمودار خوشه‌بندی مشاهدات، را به‌عنوان مراحل اساسی خوشه‌بندی در نظر گرفت. ما برای هر یک از این مراحل، در نرم‌افزارهای R، SAS و MATLAB برنامه‌ی مربوط به هر مرحله را به‌طور جداگانه نوشته‌ایم. در این مقاله با درکنار هم قرار دادن دستورهای خوشه‌بندی سلسله‌مراتبی و خوشه‌بندی K-میانگین، در نرم‌افزارهای R، SAS و MATLAB اختیار انتخاب استفاده از نرم‌افزار با توجه به نیازهای پژوهشگر ایجاد شده است.

توضیحات

1. Clustering
2. Hierarchical clustering
3. K-means clustering
4. Euclidean distance
5. Single Linkage (Nearest Neighbor)
6. Complete Linkage (Farthest Neighbor)
7. Average Linkage
8. Median
9. Centroid

مرجع‌ها

[۱] هشیارمنش، حسین (۱۳۹۲). دسته‌بندی به کمک تکنیک‌های خوشه‌بندی سلسله‌مراتبی. شار، تهران.

[2] Brock, G., Pihur, V., Datta, S. and Datta, S. (2008). clVali :An R Package for Cluster Validation. *Journal of Statistical Software*, **25**, 1-22.

[3] Cluster analysis. (2014, July 26). Wikipedia. Retrieved August 4, 2014, URL: http://en.wikipedia.org/wiki/Cluster_analysis.

[4] Deza, E. and Deza, M. (2013). *Encyclopedia of Distances*. Springer, New York.

[5] Gan, G., Ma, Ch. and Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications, Society for Industrial & Applied Mathematics*. New York.

[6] Jie, C., Zhiang, W., Junjie, W. and Wenjie, L. (2013). Towards Information Theoretic K-means Clustering for Image Indexing. Signal *Processing*, **93**, 2026–2037.

[7] Kaiyang, L. Guizhong, L., Li, X. and Chaoteng, L. (2013). A Sample Based Hierarchical Adaptive *K*-means Clustering Method for Large-scale Video Retrieval. *Knowledge-Based Systems*, **49**, 123–133.

[8] Krotha, R. and Merugula, S. (2013). A Brief Survey on Document Clustering Techniques Using MATLAB. *Journal of Computer & Organization Trends*, **3**, 1-6.

[9] Langfelder, P. and Horvath, S. (2012). Fast R Functions for Robust Correlations and Hierarchical Clustering. *Journal of Statistical Software*, **46**, 1-17.

[10] Milligan, G.W. and Cooper, M.C. (1985). An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika*, **50**, 159–179.

[11] Recchia, A. (2010). Contiguity-Constrained Hierarchical Agglomerative Clustering Using SAS. *Journal of Statistical Software*, **33**, 1-12

[12] Rencher, A. (2003). *Methods of Multivariate Analysis*. Wiley, New York.

[13] Tudor, B. (2013). Unsupervised SIFT-based Face Recognition Using an Automatic Hierarchical Agglomerative Clustering Solution. *Procedia Computer Science*, **22**, 385–394.

پیوست

بخشی از داده‌های شکل ۱ به‌صورت زیر است:

6.013469
10.389174
5.929998
11.087886
4.522113
10.649166
2.898171
9.592089
4.559451
10.467628
5.907001
10.429867
9.933838
5.294195
3.544159
6.330147
9.456396
9.789452
10.343817
6.247800
7.546953
7.180052
10.480802
5.709263
3.858021
4.591026
5.799245
3.994749
3.027266
5.001022
7.151187
3.534116
9.621552
6.444580
9.734700

حسین هشیارمنش
فوق لیسانس آمار
تهران، خیابان سید جمال‌الدین اسدآبادی، خیابان ۲۵، شماره‌ی ۵، پژوهشکده‌ی آمار.
رایانشانی: h.hoshyarmanesh@aut.ac.ir

میلاد فرهادی
فوق‌دیپلم عمران
تهران، بلوار کوهسار، مؤسسه‌ی آموزش عالی صدرا.
رایانشانی: milad_farhadi1990@yahoo.com

علی هشیارمنش
فوق لیسانس علوم دامی
کرج، مهرشهر، بلوار ارم، خیابان آزادی، دانشکده‌ی کشاورزی و منابع طبیعی دانشگاه آزاد اسلامی واحد کرج.
رایانشانی: ali.hoshyarmanesh@gmail.com

نگار جعفریان
فوق لیسانس آمار
تهران، خیابان حافظ، شماره‌ی ۴۲۴، دانشگاه صنعتی امیرکبیر.
رایانشانی: negar_jafariyan@yahoo.com