

ماتریس‌های فاصله و مقیاس‌گذاری چندبعدی

محمد باقر اسدی* و رحیم علیزاده

چکیده. مقیاس‌گذاری چندبعدی ابزاری قوی در حل مسائل مربوط به کاهش بُعد و به تصویر کشیدن داده‌ها است. در این روش با داشتن شباهت‌های بین اعضای مجموعه‌ای متناهی از اشیاء به هر عضو، مختصاتی در فضایی با بُعد پایین چنان نظیر می‌گردد که فاصله مختصات نظیرشده بین دو شیء تا حد امکان به شباهت آن‌ها نزدیک باشد. در این مقاله به معرفی ماتریس‌های فاصله و مقیاس‌گذاری چندبعدی و ذکر چند کاربرد آن‌ها می‌پردازیم.

۱. مقدمه

یادگیری ماشین^۱ و داده‌کاوی^۲ شاخه‌هایی از هوش مصنوعی‌اند که به استخراج ویژگی‌هایی از روی مجموعه‌ای از داده‌ها و طراحی یک سامانه تشخیص خودکار (یا نیمه‌خودکار) از روی آن‌ها می‌پردازند. استخراج ویژگی^۳ و کاهش بُعد^۴ از جمله روش‌هایی‌اند که علاوه بر کاهش هزینه‌های محاسباتی در روش‌های یادگیری ماشین، به‌طور مستقیم به‌صورت شیوه‌های کارآمد در داده‌کاوی به کار می‌روند. به زبان ریاضی، به هر تابع خطی یا غیرخطی از متغیرهای ورودی مسئله، ویژگی اطلاق می‌شود به‌شرطی که به بهینه‌ترین صورت ممکن پراکندگی داده‌ها را توجیه کند. به‌عنوان نمونه، اگر مجموعه داده‌های ما تصاویر صورت انسان‌ها باشد، متغیرها تعداد پیکسل‌های تصویرند که به‌طور معمول تعداد آن‌ها بسیار زیاد است و ویژگی‌های بارز، توابعی از مقادیر پیکسل‌های نقاطی از تصویر صورت مانند چشم‌ها، ابرو، بینی، دهان، گونه و نظایر آن‌ها هستند که هم بسیار کم تعدادتر از متغیرها بوده و هم تغییرات آن‌ها در بین افراد مختلف از تغییرات هر تابع دیگری از متغیرها بیشتر است. به بیان دقیق‌تر، تغییرات این ویژگی‌ها درصد زیادی از تغییرات کل داده‌ها را در تمامی پیکسل‌های تصویر به خود اختصاص می‌دهند. به همین دلیل در مسئله کاهش بُعد، یافتن تعداد بعد مناسب و تعیین صفت‌هایی که در بیان شباهت‌ها یا تفاوت‌ها نقش پررنگ‌تری دارند اهمیت ویژه‌ای دارد.

اغلب داده‌ها به‌صورت نقاطی در یک فضای اقلیدسی ثبت می‌شوند؛ در این حالت، ویژگی‌ها را می‌توان به‌صورت تصویر نقاط داده بر روی یک خمینه در فضای اقلیدسی تصور کرد که بعد خمینه کوچک‌تر از بعد فضای اقلیدسی است و این خمینه، نزدیک‌ترین خمینه به مجموعه نقاط داده با کوچک‌ترین بعد ممکن است؛ به چنین روش کاهش بُعد، یادگیری خمینه گوئیم. یادآوری می‌کنیم که هر خمینه با یک فضای اقلیدسی به‌طور موضعی همسان‌ریخت است.

عبارات و کلمات کلیدی: هوش مصنوعی، یادگیری ماشین، داده‌کاوی، ماتریس‌های فاصله، مقیاس‌گذاری چندبعدی

دبیر تخصصی رابط: علیرضا عبدالمهدی

تاریخ دریافت: ۱۴۰۰/۰۳/۲۴ تاریخ پذیرش: ۱۴۰۰/۰۶/۰۸

نوع مقاله: پژوهشی

* نویسنده مسئول

<http://dx.doi.org/10.22108/MSCI.2021.129105.1436>

¹machine learning ²data mining ³feature extraction ⁴dimensionality reduction

اگر خمینه مورد نظر، خمینه‌ای خطی، یعنی زیرفضایی خطی از فضای اقلیدسی باشد، روش کاهش بعد و استخراج ویژگی، کاهش بعد خطی یا یادگیری خمینه خطی نامیده می‌شود. معروف‌ترین و قدیمی‌ترین روش استخراج ویژگی خطی، روش تحلیل مولفه‌های اصلی^۵ است. در این روش، داده‌ها بر روی زیرفضایی خطی با کوچک‌ترین بعد ممکن به‌گونه‌ای تصویر می‌شود که تغییرپذیری (واریانس) داده‌های جدید بیشینه باشد.

با وجود این، این فرض که همواره داده‌ها حول یک خمینه خطی در فضای اقلیدسی پراکنده شده باشد، فرض دقیقی نیست. برای استخراج ویژگی در چنین حالت‌هایی، روش‌های یادگیری خمینه غیرخطی توسعه یافته‌اند. از جمله این روش‌ها می‌توان به روش مقیاس‌گذاری چندبعدی^۶، روش جانشانی خطی موضعی^۷ و روش ایزومپ^۸ اشاره کرد. در این مقاله به شرح یکی از این روش‌ها با نام مقیاس‌گذاری چندبعدی می‌پردازیم.

اغلب در فنون داده‌کاوی^۹، داده‌ها به‌صورت بردارهایی n تایی از اعداد حقیقی‌اند و بسیاری از روش‌های داده‌کاوی با این نوع داده‌ها سروکار دارند. با این حال در برخی از موارد امکان دسترسی به داده‌ها به‌صورت بردارهای n تایی دشوار، پرهزینه و زمان‌بر است و در عوض تفاوت یا شباهت بین داده‌ها به‌صورت عددی حقیقی نامنفی در دسترس است. در این‌گونه موارد اطلاعاتی که در قالب یک مجموعه داده در دسترس است، به شکل یک ماتریس قرابت^{۱۰} است. از آنجا که تفاوت یا شباهت رابطه‌ای متقارن است، ماتریس قرابت ماتریسی نامنفی و متقارن است که درایه‌های روی قطر آن همگی صفرند.

در مقیاس‌گذاری چندبعدی با داشتن یک ماتریس قرابت مربعی مانند $D = (d_{ij})$ در $\mathbb{R}^{n \times n}$ یا همان ماتریس‌های $n \times n$ با درایه‌های حقیقی و عدد طبیعی $k \geq 1$ که به‌طور معمول بسیار کوچک‌تر از n است، به دنبال یافتن نقاط $X_1, X_2, \dots, X_n$ در $\mathbb{R}^k$ هستیم که مقدار

$$\sum_{i,j} (d_{ij} - \|X_i - X_j\|_2^2)^2$$

کمینه گردد.

ماتریس $D \in \mathbb{R}^{n \times n}$ یک ماتریس فاصله اقلیدسی^{۱۱} است هرگاه نقاط $X_1, X_2, \dots, X_n \in \mathbb{R}^k$ چنان موجود باشند که

$$D = (\|X_i - X_j\|_2^2).$$

به کوچک‌ترین عدد طبیعی k که در تعریف بالا صدق کند بعد نشاننده ماتریس D گفته می‌شود. بنابراین، ارتباط تنگاتنگی میان مقیاس‌گذاری چندبعدی و ماتریس‌های فاصله اقلیدسی وجود دارد. اگر در مقیاس‌گذاری چندبعدی $k = 2, 3$ در نظر گرفته شود، داده‌ها به‌راحتی قابل بصری‌سازی در $\mathbb{R}^2$ یا $\mathbb{R}^3$ هستند که کمک زیادی به تحلیل آن‌ها خواهد کرد. برای حالات $k \geq 4$ نیز بعد داده‌ها به k کاهش یافته و با غلبه بر معضل نفرین بعد^{۱۲} با استخراج ویژگی‌های مهم‌تر و جدیدتر، تحلیل داده‌ها ساده‌تر و موثرتر خواهد شد. با توجه به اهمیت مسئله مقیاس‌گذاری چندبعدی، فنون متنوعی مانند الگوریتم‌های کلاسیک و اسماکوف^{۱۳} برای حل آن پیشنهاد گردیده است.

۲. ماتریس‌های فاصله

اگر به برخی از تقویم‌های دفتری قدیمی نگاهی انداخته باشید، شاید جدولی در انتهای آن دیده باشید که فاصله شهرهای مهم ایران از یکدیگر، در آن به تصویر کشیده شده است. حال به این بیندیشید که هیچ تصویری از نقشه ایران و

^۵principal component analysis ^۶multidimensional scaling ^۷local linear embedding ^۸ISOMAP ^۹data mining ^{۱۰}proximity matrix ^{۱۱}Euclidean distance matrix ^{۱۲}curse of dimensionality ^{۱۳}Smacof

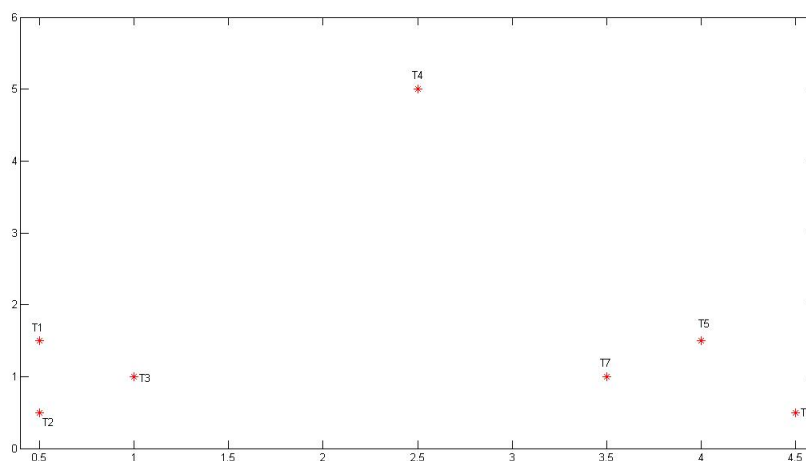
شهرهای مهم آن نداریم و فقط با کمک جدول فاصله بین شهرها بخواهیم موقعیت این شهرها را در نقشه ترسیم کنیم. آیا این کار امکان‌پذیر است؟ مقیاس‌گذاری چندبعدی به این سوال پاسخ می‌دهد.

مثالی دیگر، فرض کنید از دانشجویان دانشکده‌ای بخواهیم درباره تفاوت استادان آن دانشکده از دیدگاه خودشان اظهار نظر کنند. برای این کار از دانشجویان می‌خواهیم که عددی بین ۰ و ۵ به هر جفت از استادان نظیر کنند به‌طوری‌که اعداد کوچک‌تر بیان‌کننده تفاوت کمتر و اعداد بزرگ‌تر بیان‌کننده تفاوت بیشتر باشند. حال اگر بخواهیم با یک نگاه اجمالی به نتایج این نظرخواهی، دریابیم از نظر دانشجویان کدام استادان نسبت به هم تفاوت کمتر (یا بیشتر) دارند از چه روشی می‌توانیم استفاده کنیم؟

در این مورد می‌توان میانگین نظرات دانشجویان درباره تفاوت استادان را در یک جدول قرار داد. سپس در صفحه مختصات، به هر استاد یک نقطه را طوری نظیر کنیم که فاصله دو نقطه از هم تا حد امکان به عدد مربوط به تفاوت استادان متناظر با آن دو نقطه که در جدول آمده است، نزدیک باشد. حال به‌صورت شهودی و با یک نگاه گذرا به این نقاط، اطلاعات مفیدی درباره تفاوت بین استادان حاصل خواهد شد. به عنوان مثال، فرض کنید این نظرسنجی درباره استادان $T_1, \dots, T_7$ دانشکده انجام شده باشد و جدول میانگین نظرات دانشجویان به‌صورت زیر باشد

	T1	T2	T3	T4	T5	T6	T7
T1	0	1	0.70	4.03	3.50	4.12	3.04
T2	1	0	0.70	4.92	3.64	4.00	3.04
T3	0.70	0.70	0	4.27	3.04	3.53	2.50
T4	4.03	4.92	4.27	0	3.64	4.74	4.47
T5	3.50	3.64	3.04	3.64	0	1.12	0.70
T6	4.12	4.00	3.53	4.74	1.12	0	1.12
T7	3.04	3.04	2.50	4.47	0.70	1.12	0

در این صورت، با کمک مقیاس‌گذاری چندبعدی، موقعیت هندسی نقاط متناظر با استادان به‌صورت زیر خواهد بود



همان‌طور که مشاهده می‌کنید استادان T_1, T_2, T_3 (همچنین استادان T_5, T_6, T_7) مشابه یکدیگرند ولی استاد T_4 با بقیه تفاوت دارد.

در تعریف مقیاس‌گذاری چندبعدی علاوه بر فاصله اقلیدسی می‌توان از مترها و نرم‌های دیگر از جمله نرم یک $(\|\cdot\|_1)$ و نرم بی‌نهایت $(\|\cdot\|_\infty)$ استفاده کرد.

ماتریس $D = (d_{ij}) \in \mathbb{R}^{n \times n}$ را ماتریس فاصله نسبت به متر d گوئیم هرگاه برای یک عدد طبیعی k ، نقاط $X_1, X_2, \dots, X_n$ در $\mathbb{R}^k$ چنان یافت شوند که برای هر i و j ، $d_{ij} = d(X_i, X_j)$ ، به‌ویژه، اگر $1 \leq p \leq \infty$ ، $D = (d_{ij})$ را یک ماتریس فاصله نسبت به نرم p گوئیم اگر

$$d_{ij} = \|X_i - X_j\|_p.$$

به کوچک‌ترین عدد طبیعی k که در تعریف ماتریس فاصله صدق کند، بعد نشاننده آن گوئیم و آن را با $E(D)$ نمایش می‌دهیم. همچنین، ماتریس $X = [X_1 \ X_2 \ \dots \ X_n]^t \in \mathbb{R}^{n \times k}$ را ماتریس موقعیت ماتریس D گوئیم و می‌نویسیم $D = D(X)$.

اگر $D \in \mathbb{R}^{n \times n}$ یک ماتریس فاصله اقلیدسی باشد، رابطه‌ای بین بعد نشاننده و رتبه ماتریس D به‌صورت زیر وجود دارد [۵]:

$$\text{rank}(D) - 2 \leq E(D) \leq \text{rank}(D) - 1.$$

اگر $D \in \mathbb{R}^{n \times n}$ یک ماتریس فاصله نسبت به نرم بی‌نهایت باشد، ممکن است کران پایین رابطه بالا برقرار نباشد و امکان دارد بعد نشاننده D برابر با عدد یک باشد. در حالتی که ماتریس D معکوس‌پذیر باشد؛ یعنی $\text{rank}(D) = n$ ، کران بالای رابطه فوق برقرار است [۱۵].

ارتباط تنگاتنگی بین ماتریس‌های فاصله اقلیدسی و ماتریس‌های مثبت معین وجود دارد. با کمک ماتریس‌های مثبت معین اطلاعاتی از ماتریس‌های فاصله اقلیدسی به‌دست می‌آید که در بخش‌های بعد به آن اشاره می‌کنیم.

۱.۲. مشخص‌سازی ماتریس‌های فاصله اقلیدسی. در زیر چند مشخص‌سازی برای ماتریس‌های فاصله اقلیدسی آورده شده است. این مطالب از [۹] برگرفته شده‌اند. در موارد زیر $D = (d_{ij}) \in \mathbb{R}^{n \times n}$ یک ماتریس متقارن نامنفی است که درایه‌های روی قطر آن همگی صفرند.

(۱) ماتریس D یک ماتریس فاصله اقلیدسی است اگر و تنها اگر ماتریس

$$\hat{D} = (d_{in} + d_{jn} - d_{ij})$$

ماتریس نیمه معین مثبت باشد.

(۲) D یک ماتریس فاصله اقلیدسی است اگر و تنها اگر LDL معین مثبت باشد. در اینجا $L = I_n - \frac{1}{n}ee^t$ ، I_n ماتریس همانی و e برداری در $\mathbb{R}^n$ است که تمام درایه‌های آن یک است. رتبه این ماتریس همان بعد نشاننده D خواهد بود.

(۳) D یک ماتریس فاصله اقلیدسی است اگر و تنها اگر ماتریس $\begin{bmatrix} D & e \\ e^t & 0 \end{bmatrix}$ فقط یک مقدار ویژه مثبت داشته باشد.

(۴) ماتریس $U \in \mathbb{R}^{n \times (n-1)}$ را که در آن $V = \left(\frac{e}{\sqrt{n}} \mid U \right)$ متعامد است در نظر بگیرید. در این صورت D یک ماتریس فاصله اقلیدسی است اگر و تنها اگر $U^t D U$ یک ماتریس نیمه معین منفی باشد. رتبه این ماتریس همان بعد نشاننده D خواهد بود.

مشخص‌سازی‌های (۲) و (۴) در بالا به‌طور دقیق بعد نشاننده ماتریس‌های فاصله اقلیدسی را مشخص می‌کنند. علاوه بر این با کمک رتبه این ماتریس‌ها محدوده‌ای برای بعد نشاننده مشخص می‌شود. کران پایین این محدوده از رابطه زیر بین بعد نشاننده و رتبه ماتریس‌های فاصله اقلیدسی به‌دست می‌آید [۱۱]:

$$\text{rank}(D) \leq E(D) + 2.$$

در مشخص‌سازی (۴) دیدیم بعد نشاننده D برابر رتبه ماتریس $U^t D U$ است که یک ماتریس نیمه معین منفی و زیرماتریس $V^t D V$ است. لذا D حداقل به تعداد $E(D)$ مقدار ویژه منفی دارد و از آنجا که $\text{tr}(V^t D V) = 0$ ، $V^t D V$ حداقل یک مقدار ویژه مثبت دارد. پس D دارای حداقل $E(D) + 1$ مقدار ویژه ناصفر است. بنابراین، داریم $\text{rank}(D) = E(D) + 2$ یا $\text{rank}(D) = E(D) + 1$.

۲.۲. ماتریس‌های فاصله نسبت به نرم p . از ضرب آدامار یک ماتریس فاصله نسبت به نرم دو در خودش یک ماتریس فاصله اقلیدسی ایجاد می‌گردد. از این جهت یک مشخص‌سازی مشابه با ماتریس‌های فاصله اقلیدسی برای ماتریس‌های فاصله نسبت به نرم دو وجود دارد. در [۱۹] ثابت شده است که ماتریس $D = (d_{ij})$ یک ماتریس فاصله نسبت به نرم دو با ماتریس موقعیت $X = [X_0 \ X_1 \ \dots \ X_n]^t$ است اگر و تنها اگر ماتریس

$$(1) \quad A = \left(\frac{1}{4} [\|X_0 - X_i\|_p^2 + \|X_0 - X_j\|_p^2 - \|X_i - X_j\|_p^2] \right),$$

یک ماتریس نیمه معین مثبت باشد و بعد نشاننده D برابر رتبه ماتریس A است. از جمله تفاوت‌های ماتریس‌های فاصله اقلیدسی با ماتریس‌های فاصله نسبت به نرم دو در تعداد مقدار ویژه‌های منفی این ماتریس‌ها است. در [۱۴] ثابت شده است ماتریس‌های فاصله نسبت به نرم دو که درایه‌های خارج قطر آن همگی ناصفرند، دقیقاً یک مقدار ویژه مثبت و $n - 1$ مقدار ویژه منفی دارند.

در [۱۸] نشان داده شده است برای $n \geq 3$ حداکثر بعد نشاننده ماتریس‌های فاصله نسبت به نرم یک $\frac{n(n-1)}{4} - 1$ است. اگر $n \geq 4$ ، یک کران پایین برای بعد نشاننده ماتریس‌های فاصله نسبت به نرم یک به‌صورت $\frac{(n-2)(n-3)}{4}$ ارائه شده است [۱].

ماتریس‌های فاصله نسبت به نرم بی‌نهایت شامل تمام ماتریس‌های فاصله نسبت به هر متر دلخواهی‌اند. در حقیقت برای هر ماتریس فاصله D نسبت به هر متر دلخواه d ، نقاط $X_1, X_2, \dots, X_n \in \mathbb{R}^{n-1}$ چنان یافت می‌شوند که

$$D_{ij} = \|X_i - X_j\|_\infty.$$

برای دیدن اثبات این مطلب، فرض کنیم D یک ماتریس فاصله در $\mathbb{R}^{n \times n}$ نسبت به متر d به شکل زیر باشد

$$D = \begin{bmatrix} 0 & d_{12} & d_{13} & \dots & \dots & d_{1n} \\ d_{21} & 0 & d_{23} & \dots & \dots & d_{2n} \\ d_{31} & d_{32} & 0 & d_{34} & \dots & d_{3n} \\ \vdots & \vdots & & \ddots & & \vdots \\ d_{n1} & d_{n2} & \dots & & & 0 \end{bmatrix}.$$

حال کافی است نقاط X_i را به‌صورت زیر انتخاب کنیم

$$(2) \quad X_i = (d_{i2}, d_{i3}, \dots, d_{in}).$$

با توجه به اینکه نامساوی مثلث بین درایه‌های ماتریس D برقرار است، با محاسبه ماتریس فاصله نقاط X_i نسبت به نرم بی‌نهایت به ماتریس D می‌رسیم.

قبل از صحبت درباره بعد نشاننده این ماتریس‌ها به این نکته دقت کنید که گرچه تمام ماتریس‌های فاصله نسبت به هر متری یک ماتریس فاصله نسبت به نرم بی‌نهایت‌اند ولی لزومی ندارد که بعد نشاننده ماتریس فاصله نسبت به یک متر دلخواه با بعد نشاننده ماتریس فاصله نسبت به نرم بی‌نهایت یکسان باشد. یک کران بالا برای بعد نشاننده ماتریس‌های فاصله نسبت به نرم بی‌نهایت $n-1$ است. اگر حداکثر بعد نشاننده ماتریس $n \times n$ نسبت به نرم بی‌نهایت را با $E_{\infty}(n)$ نمایش دهیم، در [۷] برای $n \geq 4$ ثابت شده است

$$(۳) \quad \left\lfloor \frac{2n}{3} \right\rfloor \leq E_{\infty}(n) \leq n-2.$$

به راحتی می‌توان نشان داد $E_{\infty}(n+1) \leq E_{\infty}(n) + 1$. [۲۰].
همچنین در [۲۰] با ارائه چهار نقطه در $\mathbb{R}^2$ ثابت شده است حداکثر بعد نشاننده برای ماتریس‌های $n \leq 4$ نسبت به نرم بی‌نهایت برابر دو است. بعد نشاننده و رتبه ماتریس‌های فاصله نسبت به نرم بی‌نهایت در $\mathbb{R}^{n \times n}$ با درایه‌های خارج قطر ناصفر وقتی $1 < n \leq 4$ ، در رابطه زیر صدق می‌کنند

$$E(D) \leq \text{rank}(D) - 1.$$

۳.۲. تاریخچه. مطالعه ویژگی‌هایی ماتریس‌های فاصله اقلیدسی کمتر از یک قرن قدمت دارد. مفهوم فضاهای متریک و فاصله‌ها را فرشه^{۱۴} در سال ۱۹۰۶ [۳] معرفی کرد. مطالعه با جزئیات بیشتر درباره ماتریس‌های فاصله اقلیدسی را شونبرگ^{۱۵} در [۱۳] به طور عمیق‌تر در سال ۱۹۳۵ آغاز کرد. یانگ^{۱۶} و هاوس هولدر^{۱۷} در سال ۱۹۳۸ [۲۱] ویژگی‌های بیشتری را اثبات کردند و توانستند بعد نشاننده ماتریس‌های فاصله اقلیدسی را محاسبه کنند. در سال ۱۹۵۲ تاگرسن^{۱۸} مفهوم مقیاس‌گذاری چندبعدی را در قالب مسئله پیدا کردن مجموعه نقاطی که یک ماتریس فاصله را تولید می‌کند، معرفی کرد [۱۷]. ارتباط بین فاصله‌های تولیدشده و شباهت‌های اولیه غیرمتریک مبهم بود تا اینکه در سال ۱۹۶۴ کروسکال^{۱۹} مفهوم استرس را پیشنهاد کرد [۸]. تابع استرس و مینیم‌سازی آن، حوزه جدیدی در بحث مقیاس‌گذاری چندبعدی باز کرد که نمونه‌ای از آن‌ها را در [۱۲]، [۱۶] و [۱۰] می‌توان دید.
چندین مقاله درباره مسئله به دست آوردن کمترین عددی که بتوان یک فضای متریک متناهی را در فضای l_p نشاناند وجود دارد. در مقالات هاوس هولدر و یانگ دیده می‌شود این عدد برای ماتریس‌های فاصله اقلیدسی $n-1$ است. هولستنسکی^{۲۰} در سال ۱۹۷۸ [۷] برای l_{∞} ویتسن‌هاوزن^{۲۱} در سال ۱۹۸۶ [۲۰] برای l_1 و بال^{۲۲} در سال ۱۹۹۰ برای هر p به جواب این سوال پرداخته‌اند [۱]. در سال ۱۹۸۵ گاور^{۲۳} ارتباط بین رتبه ماتریس‌های فاصله اقلیدسی با بعد نشاننده را مشخص کرد [۵].

۳. چند الگوریتم حل مسئله مقیاس‌گذاری چندبعدی

در مقیاس‌گذاری چندبعدی به دنبال حل مسئله بهینه‌سازی زیر هستیم

$$\text{argmin}\{\|P - D(X)\|_F : X \text{ ماتریسی در } \mathbb{R}^{n \times k} \text{ است}\},$$

که در آن $P = (p_{ij})$ یک ماتریس قرابت و $D(X)$ ماتریس فاصله اقلیدسی متناظر با ستون‌های ماتریس X است. اگر $k=2, 3$ ، آن‌گاه داده‌ها به راحتی قابل بصری‌سازی در $\mathbb{R}^2$ یا $\mathbb{R}^3$ ‌اند. اگر p_{ij} تفاوت داده i ام با j ام باشد، در این صورت هر چه مقدار p_{ij} کوچکتر (بزرگ‌تر) باشد آنگاه در تصویر به دست آمده از مقیاس‌گذاری چندبعدی، فاصله X_i از X_j کمتر (بیشتر) خواهد بود.

¹⁴Fréchet ¹⁵Schoenberg ¹⁶Young ¹⁷Householder ¹⁸Togerson ¹⁹Kruskal ²⁰Holsztynski ²¹Witsenhausen ²²Ball ²³Gower

برای حل مسئله مقیاس‌گذاری چندبعدی روش‌های مختلفی وجود دارد؛ ولی در حالت کلی می‌توان آن‌ها را به دو دسته کلاسیک و الگوریتم تکرار^{۲۴} تقسیم کرد. روش کلاسیک که قدیمی‌ترین روش حل مسئله مقیاس‌گذاری چندبعدی است روشی جبری برای حل مسئله مقیاس‌گذاری چندبعدی است. همچنین به‌عنوان نمونه‌هایی از الگوریتم تکرار می‌توان از الگوریتم اسماکوف^{۲۵} و الگوریتم گلانت^{۲۶} نام برد [۲].

برای نمونه، با داشتن ماتریس $P = (p_{ij}) \in \mathbb{R}^{n \times n}$ ، الگوریتم کلاسیک را می‌توان به‌صورت زیر بیان کرد

$$(۱) \text{ ماتریس } A = (a_{ij}) \text{ را با درایه‌های } a_{ij} = \frac{-1}{2} p_{ij}^2 \text{ محاسبه می‌کنیم.}$$

$$(۲) \text{ ماتریس } B = (I - \frac{1}{n} ee^t) A (I - \frac{1}{n} ee^t) \text{ که در آن } I \text{ ماتریس همانی و } e \text{ برداری است که تمام درایه‌های آن یک است را به‌دست می‌آوریم.}$$

$$(۳) \text{ تجزیه مقدار ویژه } B^{۲۷} \text{ به‌شکل } Q \Lambda Q^t \text{ که در آن } Q \text{ ماتریس متعامد و } \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) \text{ ماتریسی قطری است با شرط } \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \text{ را به‌دست آورده و تمام مقادیر ویژه منفی } \Lambda \text{ را برابر صفر قرار می‌دهیم.}$$

$$(۴) \text{ فرض کنیم بخواهیم مقیاس‌گذاری چندبعدی را در } k \text{ بُعد حل کنیم. ماتریس } \Lambda_k \text{ حاصل از } k \text{ سطر و ستون اول } \Lambda \text{ و ماتریس } Q_k \text{ حاصل از } n \text{ سطر و } k \text{ ستون اول } Q \text{ را در نظر گرفته و ماتریس } X = Q_k \Lambda_k^{1/2} \text{ را به‌دست می‌آوریم که در آن } X_1, X_2, \dots, X_n \text{ به ترتیب سطرهای اول تا } n \text{ ام ماتریس } X \text{ است.}$$

۴. کاربردهایی از مقیاس‌گذاری چندبعدی

کاربردهای مقیاس‌گذاری چندبعدی را در علوم مختلف می‌توان دید. در اینجا به چند مورد اشاره خواهیم کرد.

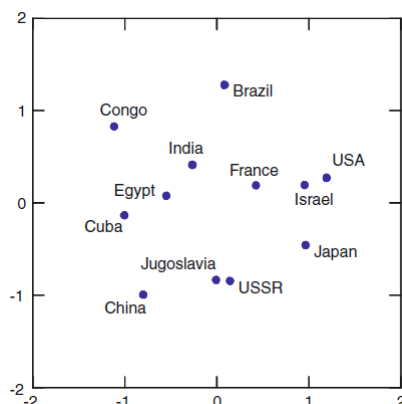
۱.۴. روش ویش در استخراج معیارهای دسته‌بندی کشورها. یکی از مباحث روانشناسی یافتن دلیل شباهت‌هایی است که در ذهن افراد وجود دارد. روش‌های حل مقیاس‌گذاری چندبعدی در این مباحث ممکن است بسیار مفید باشد. در چنین مدل‌هایی اشیاء مورد قضاوت به‌صورت نقاطی در صفحه در نظر گرفته می‌شوند و هر چه فاصله دو نقطه بیشتر باشد نشان‌دهنده این است که بین دو شیء متناظر به آن دو نقطه، شباهت کمتری وجود دارد. سختی تحلیل این مدل‌ها در این است که محقق نمی‌داند افراد چه معیارهایی را برای ارزیابی انتخاب کرده‌اند.

نمونه‌ای از این مدل‌ها مثالی است که ویش^{۲۸} در سال ۱۹۷۱ در نظر گرفت [۲]. او به دنبال پاسخگویی به این سوال بود که چه دلایلی باعث می‌شود افراد، کشورهای مختلف را شبیه به هم یا متفاوت از هم ارزیابی کنند. برای این کار او از ۱۸ دانش‌آموز خواست تا ۱۲ کشور را ارزیابی کنند و از هریک خواست که هر دو زوج از کشورها را با اعداد ۱ تا ۹ که ۱ بیانگر کمترین شباهت و ۹ نشان‌دهنده بیشترین شباهت است، رتبه‌بندی کنند. در جدول زیر میانگین پاسخ این ۱۸ دانش‌آموز را مشاهده می‌کنیم

^{۲۴}iterative MDS algorithm ^{۲۵}Smacof ^{۲۶}Glunt ^{۲۷}eigen decomposition ^{۲۸}Wish

Country		1	2	3	4	5	6	7	8	9	10	11
Brazil	1	-										
Congo	2	4.83	-									
Cuba	3	5.28	4.56	-								
Egypt	4	3.44	5.00	5.17	-							
France	5	4.72	4.00	4.11	4.78	-						
India	6	4.50	4.83	4.00	5.83	3.44	-					
Israel	7	3.83	3.33	3.61	4.67	4.00	4.11	-				
Japan	8	3.50	3.39	2.94	3.83	4.22	4.50	4.83	-			
China	9	2.39	4.00	5.50	4.39	3.67	4.11	3.00	4.17	-		
USSR	10	3.06	3.39	5.44	4.39	5.06	4.50	4.17	4.61	5.72	-	
USA	11	5.39	2.39	3.17	3.33	5.94	4.28	5.94	6.06	2.56	5.00	-
Jugoslavia	12	3.17	3.50	5.11	4.28	4.72	4.00	4.44	4.28	5.06	6.67	3.56

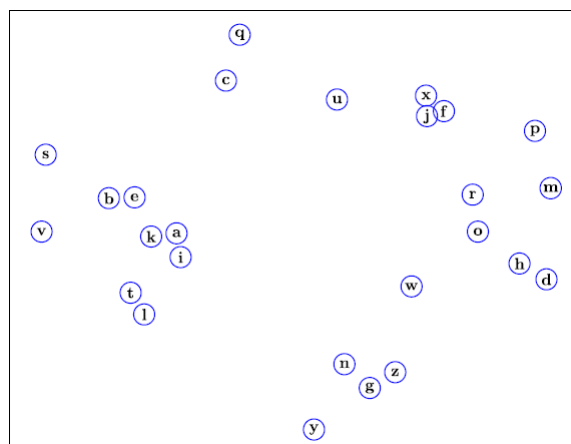
مقیاس‌گذاری چندبعدی با در نظر گرفتن بعد $k = 2$ برای این مسئله، به شکل زیر به دست می‌آید



از آنجا که با چرخش تصویر فاصله تغییری نمی‌کند، می‌توان تصویر را به اندازه‌ای چرخاند که بهتر بتوان آن‌ها را از نظر روانشناسی تحلیل کرد. ویش پیشنهاد کرد چرخش تصویر به اندازه ۴۵ درجه، تحلیل روانشناسی بهتری به ما می‌دهد. او متوجه شد که دو عامل «توسعه‌یافته بودن (نبودن)» و «کمونیست بودن (نبودن)» کشورها در انتخاب افراد تاثیر بسیاری داشته است. به عنوان مثال، کشورهای چین، یوگسلاوی، و شوروی کشورهای توسعه‌یافته کمونیست نزدیک به هم قرار گرفتند، در حالی که کشورهای آمریکا، فرانسه، و ژاپن کشورهای توسعه‌یافته غیرکمونیست کنار هم قرار گرفتند.

۲.۴. طراحی صفحه کلید. یکی از کاربردهای مقیاس‌گذاری چندبعدی را می‌توان در طراحی صفحه کلید مشاهده کرد [۱۱]. مهم‌ترین هدف در طراحی صفحه کلید این است که کلیدها در موقعیت‌هایی قرار گیرند که تایپ آن‌ها به ساده‌ترین شکل ممکن باشد.

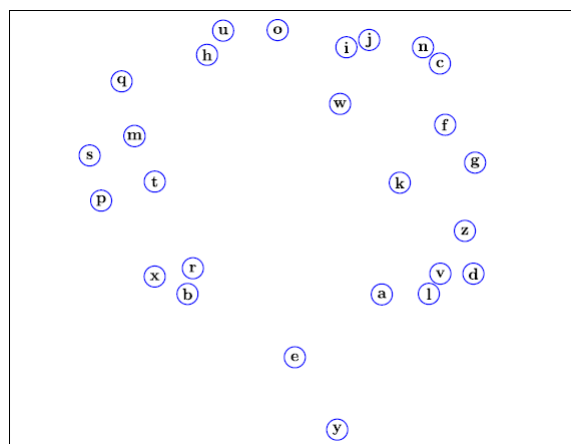
فرض کنیم فقط با یک دست بتوانیم تایپ کنیم. بهترین حالت قرارگرفتن کلیدها در صفحه کلید این است که اگر دو حرف انگلیسی در بیشتر کلمات نزدیک به هم قرار دارند، آن دو حرف در صفحه کلید نیز به هم نزدیک باشند تا مجبور نباشیم برای تایپ دو حرف نزدیک به هم فاصله بیشتری را با انگشتان دست خود طی کنیم. برای حل مسئله با استفاده از مقیاس‌گذاری چندبعدی به یک ماتریس قرابت $\hat{D}$ نیاز داریم. برای این منظور فاصله هر دو حرف انگلیسی را در هر لغت انگلیسی جمع یک با تعداد حروف موجود بین آن‌ها تعریف می‌کنیم. به عنوان مثال در لغت *love* فاصله l از o برابر یک، فاصله l از v برابر دو و فاصله l از e برابر سه است. به همین ترتیب بقیه فاصله‌ها رو در این لغت و ۷۰۰۰۰ لغت پرکاربرد انگلیسی محاسبه می‌کنیم. میانگین تمام فواصل به دست آمده برای هر دو حرف را محاسبه می‌کنیم و آن‌ها را در درایه‌های ماتریس قرابت $\hat{D}$ قرار می‌دهیم. مقیاس‌گذاری چندبعدی با در نظر گرفتن بعد $k = 2$ برای این مسئله، به شکل زیر دست می‌یابد



می‌بینیم شکل بالا ایراداتی دارد. مثلاً ترکیب حروف j و x در حروف انگلیسی به‌ندرت به کار می‌رود در حالی که در شکل بالا به هم نزدیک هستند. دلیل به وجود آمدن این اشکال در نظر نگرفتن تعداد تکرار لغات در کلمات است. برای این منظور، ماتریس $W = (w_{ij})_{32 \times 32}$ را محاسبه می‌کنیم که در آن درایه‌های این ماتریس، به‌صورت نسبت تعداد تکرارهای زوج حروف در لغات به تعداد کل لغات است. به‌طور مثال، اگر دو حرف s و l در ۳۰۰۰ از این ۷۰۰۰۰ لغت تکرار شده باشد، درایه (l, s) ماتریس W را $\frac{3}{70}$ قرار می‌دهیم. در نهایت درایه‌ها را به شکلی که بزرگ‌ترین آن برابر یک باشد نرمال می‌کنیم. حال مقیاس‌گذاری چندبعدی با مینیمم‌سازی تابع

$$\sum_i \sum_j W_{ij} (d_{ij}^Y - \hat{D}_{ij})$$

در $\mathbb{R}^2$ به شکل زیر می‌انجامد

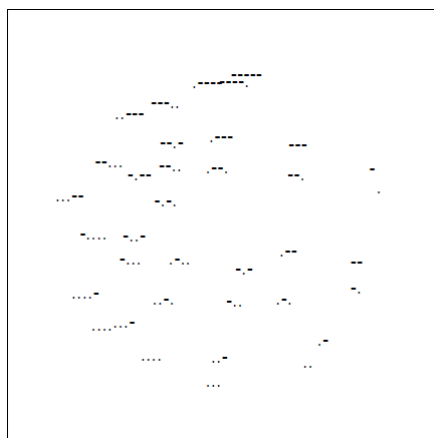


۳.۴. روانشناسی سیگنال‌های مورس. یک کاربرد دیگر مقیاس‌گذاری چندبعدی را می‌توان در سیگنال‌های مورس [۶] مشاهده کرد. محققان در اینجا به دنبال عوامل روانشناسی هستند که سبب می‌شوند افراد با شنیدن سیگنال‌های مورس دچار اشتباه شوند. برای این منظور محققان ۳۶ سیگنال را که ۲۶ تا از آن‌ها مربوط به حروف $A, B, \dots, Z$ و ۱۰ تا از آن‌ها مربوط به اعداد ۰، ۱، ...، ۹ است به‌صورت جفت‌جفت در نظر می‌گیرند و از افرادی که با این سیگنال‌ها آشنایی ندارند می‌خواهند این جفت سیگنال‌ها را گوش دهند و به این سوال پاسخ دهند که آیا این سیگنال‌ها مثل هم بوده‌اند یا خیر؟ برای این منظور ۵۹۸ فرد، ۳۵۱ جفت از سیگنال‌های مورس را قضاوت کردند و در نهایت ماتریس

قربان را بر اساس اینکه چه تعدادی دو سیگنال را شبیه به هم دانسته‌اند به دست آوردند. نتایج این تحقیقات را در جدول زیر می‌توان مشاهده کرد

Morse Code	Sign	A	B	C	D	...	0
..	A	92	4	6	13	...	3
... ..	B	5	84	37	31	...	4
... ..	C	4	38	87	17	...	12
... ..	D	8	62	17	88	...	6
...	:	:	:	:	:	...	:
-----	0	9	3	11	2	...	94

مقیاس‌گذاری چندبعدی با در نظر گرفتن بعد $k = 2$ برای این مسئله، به شکل زیر می‌انجامد



۵. موضوعات پژوهشی در مقیاس‌گذاری چندبعدی

هرچند مطالعات بر روی نظریه و روش‌های اجرای مقیاس‌گذاری چندبعدی در این چند دهه پیشرفت‌های زیادی داشته است، موضوعات متنوعی برای توسعه بیشتر این روش‌ها پیش روی پژوهشگران قرار دارد. در زیر به چند مورد اشاره می‌کنیم.

- (۱) چنانکه اشاره شد ارتباط بین بعد نشاننده ماتریس‌های فاصله اقلیدسی و مرتبه آن‌ها به طور کامل مشخص شده است. بررسی این ارتباط برای ماتریس‌های فاصله نسبت به مترهای مختلف موضوع با اهمیتی است.
- (۲) اکثر روش‌های پیاده‌سازی برای مقیاس‌گذاری چندبعدی نسبت به نرم اقلیدسی و به صورت متمرکز است. ارائه روش‌های توزیع شده برای طیف وسیع‌تری از نرم‌ها، از جمله مسائل مهم و کاربردی در این زمینه است.
- (۳) الگوریتم‌های پیاده‌سازی موجود برای ماتریس‌های با ابعاد بسیار بزرگ با دشواری‌های محاسباتی فراوانی همراه هستند. ایجاد و توسعه روش‌های مقیاس‌پذیر برای حل مسئله مقیاس‌گذاری چندبعدی از اولویت‌های تحقیقاتی در این زمینه است.

مراجع

- [1] K. Ball, Isometric embedding in l_p -spaces, *European J. Combin.*, **11** (1990) 305-311.

- [2] I. Borg, P. J. Groenen and P. Mair, *Applied Multidimensional Scaling*, Springer Science & Business Media, New York, 2012.
- [3] M. M. Fréchet, Sur quelques points du calcul fonctionnel, *Rend. Circ. Mat. Palermo*, **22** (1906) 1–72.
- [4] W. Glunt, T. L. Hayden, S. Hong and J. Wells, An alternating projection algorithm for computing the nearest Euclidean distance matrix, *SIAM J. Matrix Anal. Appl.*, **11** (1990) 589–600.
- [5] J. C. Gower, Properties of Euclidean and non-Euclidean distance matrices, *Linear Algebra Appl.*, **67** (1985) 81–97.
- [6] Patrick J. F. Groenen and I. Borg, *Past, present, and future of multidimensional scaling*, in Visualization and Verbalization of Data, J. Blasius and M. Greenacre, eds., Chapman and Hall/CRC, London, 2014, 95–117.
- [7] W. Holsztynski, $\mathbb{R}^n$ as universal metric space, *Notices Amer. Math. Soc.*, **25** (1978) A–367.
- [8] J. B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika*, **29** (1964) 1–27.
- [9] C. K. Li, T. Milligan and M. Trosset, Euclidean and circum-Euclidean distance matrices: characterizations and linear preservers, *Electron. J. Linear Algebra*, **20** (2010) 739–752.
- [10] J. J. Meulman, The integration of multidimensional scaling and multivariate analysis with optimal transformations, *Psychometrika*, **57** (1992) 539–565.
- [11] R. Parhizkar, *Euclidean Distance Matrices: Properties, Algorithms and Applications*, School of Computer and Communication Sciences, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland, 2013.
- [12] J. W. Sammon, A nonlinear mapping for data structure analysis, *IEEE Trans. Comput.*, **100** (1969) 401–409.
- [13] I. J. Schoenberg, Remarks to Maurice Fréchet's article "Sur La Definition Axiomatique D'Une Classe D'Espace Distances Vectoriellement Applicable Sur L'Espace De Hilbert", *Ann. of Math.*, **36** (1935) 724–732.
- [14] I. J. Schoenberg, On certain metric spaces arising from Euclidean spaces by a change of metric and their imbedding in Hilbert space, *Ann. of Math.*, **38** (1937) 787–793.
- [15] I. J. Schoenberg, Metric spaces and positive definite functions, *Trans. Amer. Math. Soc.*, **44** (1938) 522–536.
- [16] Y. Takane, F. W. Young and J. De Leeuw, Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features, *Psychometrika*, **42** (1977) 7–67.
- [17] W. S. Torgerson, Multidimensional scaling: I., theory and method, *Psychometrika*, **17** (1952) 401–419.
- [18] B. van Cutsem, *Classification and Dissimilarity Analysis*, Springer-Verlag, New York, 1994.
- [19] J. H. Wells and L. R. Williams, *Embeddings and Extensions in Analysis*, Ergebnisse, Springer-Verlag, Berlin, 1975.
- [20] H. S. Witsenhausen, Minimum dimension embedding of finite metric spaces, *J. Combin. Theory Ser. A*, **42** (1986) 184–199.
- [21] G. Young and A. S. Householder, Discussion of a set of points in terms of their mutual distances, *Psychometrika*, **3** (1938) 19–22.

محمد باقر اسدی

دانشکده ریاضی آمار و علوم کامپیوتر، دانشکده‌گان علوم، دانشگاه تهران، تهران

mb.asadi@ut.ac.ir

محمد باقر اسدی عضو هیئت علمی دانشگاه تهران است. وی دوره کارشناسی ریاضی (گرایش دبیری) را در دانشگاه سمنان، دوره کارشناسی ارشد ریاضی محض را در دانشگاه تهران و دوره دکتری خود را در رشته ریاضی محض (گرایش آنالیز) در دانشگاه خوارزمی گذرانده است.



رحیم علیزاده

گروه ریاضی، دانشکده علوم پایه، دانشگاه شاهد، تهران

alizadeh@shahed.ac.ir

رحیم علیزاده عضو هیئت علمی دانشگاه شاهد است. وی دوره کارشناسی ریاضی (گرایش کاربرد در کامپیوتر) و دوره‌های کارشناسی ارشد و دکتری خود را در رشته ریاضی محض (گرایش آنالیز) در دانشگاه امیرکبیر گذرانده است.

