

کاربرد الگوریتم PCA در سنجش از دور با استفاده از زبان برنامه‌نویسی پایتون (مورد مطالعاتی: تصاویر مولتی اسپکترال ماهواره لندست 9)

مهدی فرمehنی فراهانی¹، دکتر خلیل رضایی²، شیدا کرم‌الهی³، فاطمه جعفری مقدم³

¹دانشجوی کارشناسی زمین‌شناسی، دانشگاه خوارزمی

²دانشیار رسوب‌شناسی دانشکده علوم‌زمین، دانشگاه خوارزمی تهران

³دانشجوی کارشناسی ارشد رسوب‌شناسی و سنگ‌شناسی رسوبی، دانشگاه خوارزمی

*ایمیل نویسنده مسئول: khalil.rezaei@khu.ac.ir

چکیده: با وجود پیشرفت روزافزون فناوری و همه گیر شدن برنامه‌نویسی در شاخه‌های مختلف علوم پایه، الزام استفاده و یادگیری برنامه نویسی در شاخه های مختلف علوم زمین امری غیر قابل انکار است. یکی از موارد مورد توجه دنیای فناوری امروز، مبحث علوم داده و استفاده از روش‌های آماری پیشرفته با کمک برنامه‌نویسی است و با وجود توسعه بسته‌های مختلف در بستر زبان برنامه نویسی پایتون انجام کارهای سنجش از دور برای متخصصان این حوزه راحت‌تر از پیش شده است. در این مطالعه تلاش بر آن شده است که با استفاده از الگوریتم یادگیری ماشین PCA و به کمک تصاویر مولتی اسپکترال لندست 9 نویز و اشکالات تصاویر را حذف کنیم و داده‌های اصلی را استخراج کنیم.

واژگان کلیدی: جبر خطی، سنجش از دور، یادگیری ماشین، علوم داده، تحلیل مؤلفه‌های اصلی

1- مقدمه

امروزه اهمیت علوم زمین بدلیل افزایش نیاز صنایع و تکنولوژی های مختلف به مواد اولیه از یکسو و نیز مخاطرات گوناگون طبیعی برکسی پوشیده نیست. همچنین استفاده از هوش مصنوعی و تمام شاخه های آن از جمله یادگیری ماشین نیز با سرعت قابل توجهی در حال گسترش است. این بکارگیری شامل زمینه های مختلف از جمله پیش بینی رفتار آتی مواد طبیعی [7]، انجام کارها با دقت اما بدون نظارت [7]، وقوع یا عدم وقوع مخاطرات یا پیش بینی تقریبی زمان آنها (رفرنس)، کمک به جلوگیری از افزایش خسارات ناشی از مخاطرات طبیعی [17]، کمک به اکتشاف و تشخیص زون-های حاوی مواد معدنی و یا نفت و گاز [10] می باشد. اگرچه برای استفاده از توانایی‌های یادگیری ماشین نیاز به تسلط بر مباحث جبر خطی، حسابان و آمار و احتمالات دارد ولیکن با کمک و بهره‌گیری از ماژول‌های موجود در زبان‌های برنامه‌نویسی (برای مثال پایتون) نیاز خود را تا حدودی بر طرف کرد.

۲- روش و ابزار

2.1.1 یادگیری ماشین

یادگیری ماشین مبحثی است که بدون برنامه ریزی مستقیم (نظارت) به کامپیوترها قابلیت یادگیری می دهد. [1] در یادگیری بدون نظارت به علت کمبود اطلاعات موجود ماشین به دنبال اشتراکات است که با کمک آن ها بتواند داده ها را تحلیل کند، برآورد تراکم انجام دهد و در نهایت ابعاد (ویژگی ها) را کاهش دهد. [2, 3] در خوشه بندی، گروه بندی مجموعه ای از اشیاء انجام می شه. این کار به این صورت است که اشیائی که در یک خوشه قرار دارند، در مقایسه با اشیاء دیگر خوشه ها مشابه تر هستند. این یک روش معمول برای تجزیه و تحلیل داده های آماری می باشد. خوشه ها شامل گروه هایی با فاصله های کم بین اعضای خوشه، مناطق متراکم فضای داده، فواصل و یا توزیع های آماری خاص هستند. در نتیجه خوشه بندی می تواند به عنوان یک مسئله بهینه سازی چند هدفه صورت بگیرد. انتخاب الگوریتم خوشه بندی مناسب و تنظیمات پارامتر (از جمله پارامترهایی مانند تابع فاصله مورد استفاده، آستانه تراکم یا تعداد خوشه مورد انتظار) بستگی به تنظیم مجموعه داده ها توسط فرد و استفاده خاص فرد از نتایج دارد. تجزیه و تحلیل خوشه ای یک روش اتوماتیک نیست، بلکه یک فرآیند تکراری از کشف دانش یا بهینه سازی چند هدفه تعاملی هست که در اون آزمایش و شکست وجود دارد. توی خوشه بندی اغلب لازم است که داده ها پیش پردازش بشوند و پارامترهای مدل اصلاح بشوند تا نتیجه حاصل، همان نتیجه ای دلخواه باشد. [4, 5, 6]

2.1.2 تحلیل مؤلفه اصلی

تحلیل مؤلفه های اصلی یا PCA یکی از مباحث مهم تبدیلی در فضای برداری در جبر خطی است که اولین بار در سال 1901 توسط کارل پیرسون ارائه شد. [19] در روش PCA همان گونه که از نامش مشخص است با استفاده از روش های آماری سبب کاهش بعد مجموعه داده می شود که این امر منجر بر استخراج مؤلفه های اصلی با نگه داشتن حداکثر اطلاعات میشود تا یک سری ویژگی را که از اهمیت به نسبت بالاتری برخوردار است را آشکار سازد. البته لازم به ذکر است که در همی موارد منجر به حذف نمی شود و گاهی به آشکار شدن یک سری ویژگی جدید می انجامد. [18] در اصل منظور از مؤلفه های اصلی مجموعه از نقاط بر روی صفحه مختصات حقیقی است که در اصل دنباله ای از بردارهای واحد m است که در آن ها بردار n امتداد خطی است که بهترین تطابق را با داده هایمان دارد. اگر چه بیشتر در زمینه پردازش سیگنال بهینه است، ولیکن در زمینه کاهش اشکالات تصویر در فرایند پردازش تصویر می تواند مفید واقع شود. بیشترین میزان داده در مؤلفه اول است. برای پیدا کردن مؤلفه اول نیاز است که واریانس را بیشینه کنیم و اولین بردار وزن در قالب ماتریسی باید مطابق شرط زیر باشد [20]:

$$w_{(1)} = \arg \max_{\|w\|=1} \{ \|Xw\|^2 \} = \arg \max_{\|w\|=1} \{ w^T X^T X w \}$$

(1)

و از آن جایی که در جبر خطی، به حاصل تقسیم صورت درجه دوم x^*Ax و ضرب داخلی بردار مفروض در خودش x^*x خارج قسمت رابلی گفته می شود [21]، داریم که:

$$r(x) = \frac{x^* Ax}{x^* x} \quad (2)$$

$$w_{(1)} = \arg \max \left\{ \frac{w^T X^T X w}{w^T w} \right\} \quad (3)$$

در نتیجه بر اساس معادله 3 متوجه می شویم که کمیتی باید بیشینه شود را می توان خارج قسمت رایی دانست. مؤلفه های بعدی را نیز می توان از معادله 4 به دست آورد.

$$\hat{X}_k = X - \sum_{s=1}^{k-1} X w_{(s)} w_{(s)}^T \quad (4)$$

که اگر بردار وزنی که حداکثر واریانس را از این ماتریس داده خارج کنیم داریم که:

$$w_{(k)} = \arg \max_{\|w\|=1} \{ \|\hat{X}_k w\|^2 \} = \arg \max \left\{ \frac{w^T \hat{X}_k^T \hat{X}_k w}{w^T w} \right\} \quad (5)$$

2.1.3 نرمال سازی Min-Max

زمانی که می خواهیم مجموعه داده ای را به کامپیوتر فراخوانی کنیم در مرحله ی پیش پردازش باید با بهره گیری از روش های آماری آن ها را نرمال سازی کنیم. در حالی که روش های متعددی برای نرمال سازی داده ها وجود دارد ولیکن یکی از مرسوم ترین آن ها روش Min-Max است. این روش سادترین روش تغییر بازه به $[0,1]$ با بهره گیری از روش های تبدیل خطی با حفظ ارتباط و ساختار اولیه است. [22,23,24] به طور کلی روش نرمال سازی Min-Max بصورت زیر تعریف می شود:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (6)$$

2.2 پردازش تصویر

به مجموعه فرایندهایی که بر روی تصویر به منظور بهبود یا دستکاری تصویر صورت میگیرد پردازش تصویر میگویند. [11] پردازش تصویر فرآیندیست که به صورت پیکسل به پیکسل انجام می شود و با اعمال مجموعه ای از توابع (که با توجه به نوع فعالیت متفاوت است) روی هر پیکسل صورت می گیرد [11] در فرایندهای پردازش تصویر غالباً از الگوریتم های یادگیری ماشین استفاده می شود. [12]

هشتمین همایش ملی انجمن رسوب شناسی ایران



8th Symposium of Sedimentological Society of Iran / 31 Jan - 1 Feb 2024
University of Hormozgan

۱۱ و ۱۲ بهمن ماه ۱۴۰۲
دانشگاه هرمزگان

ماهواره لندست 9 یکی ماهواره های ناسا می باشد که در تاریخ 27 سپتامبر 2021 پرتاب شد و به فعالیت های رصد زمین مشغول شد. لندست 9 دارای دو سری سنجنده است: تصویربرداری زمین عملیاتی (OLI-2) و حسگر حرارتی مادون قرمز (TIRS-2) که در جدول زیر به عملکرد و مشخصات آن ها اشاره شده است. [8, 9]

#	Spectral band	Req min lower band edge (nm)	Req max upper band edge (nm)	OLI lower (nm)	OLI upper (nm)	OLI-2 lower (nm)	OLI-2 upper (nm)
1	Coastal Aerosol	433	453	435	451	435	450
2	Blue	450	515	452	512	452	512
3	Green	525	600	533	590	532	589
4	Red	630	680	636	673	636	672
5	NIR	845	885	851	879	850	879
6	SWIR 1	1560	1660	1566	1651	1565	1651
7	SWIR 2	2100	2300	2107	2294	2105	2294
8	Panchromatic	500	680	504	676	503	675
9	Cirrus	1360	1390	1363	1384	1363	1384

جدول 1 :

مشخصات و عملکرد سنسورهای ماهواره لندست 9 [8]

2.4 پایتون

پایتون که یک زبان برنامه نویسی ماژولار و سطح بالا است. [7] ماژولار بودن آن به ما کمک می کند تا بتوانیم از کتابخانه های موجود و توسعه داده شده در آن استفاده کنیم. در این مطالعه تلاش بر آن شده است تا با بهره گیری از بسته های Matplotlib, NumPy, scikit-image و scikit-learn فرایندهای محاسبات علمی، پردازش تصویر و یادگیری ماشین برای سنجش از دور صورت بگیرد. [13, 14, 15, 16]

3- نتایج و بحث

در این مطالعه با استفاده از تصاویر چند طیفی تصاویر باندهای 2، 3، 4، 5، 6 و 7 (آبی، سبز، قرمز، نزدیک فرو سرخ و فرو سرخ با طول موج کوتاه 1 و 2) ماهواره لندست 9 استفاده شده است. در ابتدا تصاویر را با استفاده از اسکرپیت زیر در محیط Jupyter Lab فراخوانی می کنیم.

هشتمین همایش ملی انجمن رسوب شناسی ایران



8th Symposium of Sedimentological
Society of Iran / 31 Jan - 1 Feb 2024
University of Hormozgan

۱۱ و ۱۲ بهمن ماه ۱۴۰۲
دانشگاه هرمزگان

```
import matplotlib.pyplot as plt
import numpy as np
from skimage import io
import skimage as skm
from sklearn.decomposition import PCA
```

```
ic = io.imread_collection('C:/Remote Sensing/data/Torud/3/LANDSAT-9/*.tif')
images = []
B2 = ic[1].shape
B3 = ic[2].shape
B4 = ic[3].shape
B5 = ic[4].shape
B6 = ic[5].shape
B7 = ic[6].shape
```

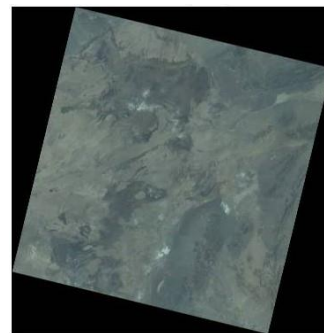
سپس با استفاده از اسکرپیت زیر تصاویر را با استفاده از روش Min-Max نرمالایز می‌کنیم.

```
NB2 = (B2-np.min(B2)) / (np.max(B2)-np.min(B2))
NB3 = (B3-np.min(B3)) / (np.max(B3)-np.min(B3))
NB4 = (B4-np.min(B4)) / (np.max(B4)-np.min(B4))
NB5 = (B5-np.min(B5)) / (np.max(B5)-np.min(B5))
NB6 = (B6-np.min(B6)) / (np.max(B6)-np.min(B6))
NB7 = (B7-np.min(B7)) / (np.max(B7)-np.min(B7))
```

با استفاده از اسکرپیت زیر تصویر اولیه را نمایش می‌دهیم.

```
rgb = np.stack([NB4,NB3,NB2], axis=2)

plt.figure(figsize=(15,10))
plt.imshow(rgb)
plt.axis('off')
plt.show()
```



شکل ۱: تصویر RGB حاصل شده از ترکیب باندهای قرمز، سبز و آبی مربوط به منطقه طرود

در ادامه با استفاده از ماژول PCA موجود در کتابخانه scikit-learn، الگوریتم را بر روی تصاویر خود اعمال می‌کنیم.

```
for i in range(len(ic)):
    images.append(ic[i].flatten())

pca = PCA()
```

هشتمین همایش ملی انجمن رسوب شناسی ایران



8th Symposium of Sedimentological
Society of Iran / 31 Jan - 1 Feb 2024
University of Hormozgan

۱۱ و ۱۲ بهمن ماه ۱۴۰۲
دانشگاه هرمزگان

```
pca.fit(images)
```

```
pc1 = pca.components_[0].reshape(shape1)  
pc2 = pca.components_[1].reshape(shape1)  
pc3 = pca.components_[2].reshape(shape1)  
pc4 = pca.components_[3].reshape(shape1)  
pc5 = pca.components_[4].reshape(shape1)  
pc6 = pca.components_[5].reshape(shape1)
```

برای نمایش تصاویر حاصل شده نیز از اسکریپت زیر استفاده می‌کنیم.

```
fig , ax = plt.subplots(nrows=6 , ncols=2, figsize=(20,30))  
ax[0,0].imshow(pc1, cmap='gray')  
ax[0,0].set_title('PCA Band 1')  
ax[0,0].axis('off')  
ax[0,1].hist(pc1.ravel(), bins=256, density=True, histtype='bar', color='black')  
ax[0,1].set_title('Histogram of PCA Band 1')  
#####  
ax[1,0].imshow(pc2, cmap='gray')  
ax[1,0].set_title('PCA Band 2')  
ax[1,0].axis('off')  
ax[1,1].hist(pc2.ravel(), bins=256, density=True, histtype='bar', color='black')  
ax[1,1].set_title('Histogram of PCA Band 2')  
#####  
ax[2,0].imshow(pc3, cmap='gray')  
ax[2,0].set_title('PCA Band 3')  
ax[2,0].axis('off')  
ax[2,1].hist(pc3.ravel(), bins=256, density=True, histtype='bar', color='black')  
ax[2,1].set_title('Histogram of PCA Band 3')  
#####  
ax[3,0].imshow(pc4, cmap='gray')  
ax[3,0].set_title('PCA Band 4')  
ax[3,0].axis('off')  
ax[3,1].hist(pc4.ravel(), bins=256, density=True, histtype='bar', color='black')  
ax[3,1].set_title('Histogram of PCA Band 4')  
#####  
ax[4,0].imshow(pc5, cmap='gray')  
ax[4,0].set_title('PCA Band 5')  
ax[4,0].axis('off')  
ax[4,1].hist(pc5.ravel(), bins=256, density=True, histtype='bar', color='black')  
ax[4,1].set_title('Histogram of PCA Band 5')  
#####  
ax[5,0].imshow(pc6, cmap='gray')  
ax[5,0].set_title('PCA Band 6')  
ax[5,0].axis('off')  
ax[5,1].hist(pc6.ravel(), bins=256, density=True, histtype='bar', color='black')
```

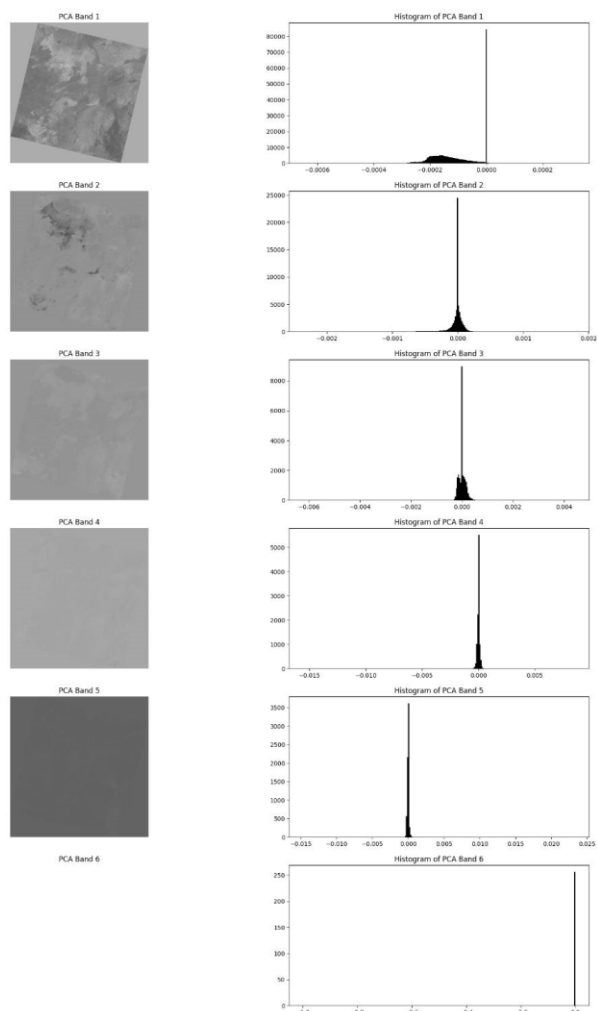
هشتمین همایش ملی انجمن رسوب شناسی ایران



8th Symposium of Sedimentological Society of Iran / 31 Jan - 1 Feb 2024
University of Hormozgan

۱۱ و ۱۲ بهمن ماه ۱۴۰۲
دانشگاه هرمزگان

```
ax[5,1].set_title('Histogram of PCA Band 6')
plt.show()
```



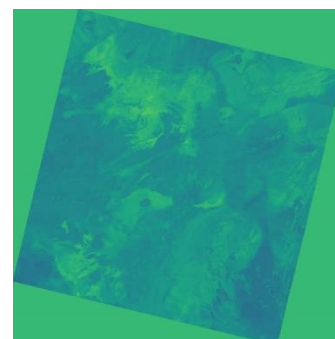
شکل (2): تصاویر مؤلفه‌های اول تا ششم حاصل شده از اعمال تحلیل مؤلفه اصلی.

هشتمین همایش ملی انجمن رسوب شناسی ایران



8th Symposium of Sedimentological
Society of Iran / 31 Jan - 1 Feb 2024
University of Hormozgan

۱۱ و ۱۲ بهمن ماه ۱۴۰۲
دانشگاه هرمزگان



شکل 3 : مؤلفه اصلی اول (PC1)

4- نتیجه گیری

اگرچه کاربرد اصلی PCA در پردازش سیگنال است ولیکن در مطالعه صورت گرفته توانسته عملکرد خوبی را از خود در پردازش تصویر از خود نشان دهد. در شکل (2) دیده می شود داده ها در یک فضای شش بعدی پردازش شده اند و بخش عمده ای آن (نزدیک به 90%) در PC1 (شکل 3) وجود دارد، و باقی داده ها در پنج مؤلفه بعدی پخش شده اند. بصورت کلی می توان گفت عمده داده های مهم در دو بعد ابتدایی قرار گرفته اند، بعد سوم و چهارم دارای جزئیات هستند و بعد پنجم و ششم اکثر نویزها و داده های کم اهمیت تر را شامل می شود. در کل می توان از این روش برای حذف نویزهای موجود در تصاویر ماهواره ای برای انجام هرچه دقیق تر فرایندهای سنجش از دور استفاده کرد.

5- مراجع

1. Samuel, Arthur L. "Machine learning." *The Technology Review* 62, no. 1 (1959): 42-45
2. Dike, H. U., Zhou, Y., Deveerasetty, K. K., & Wu, Q. (2018, October). Unsupervised learning based on artificial neural network: A review. In *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)* (pp. 322-327). IEEE.
3. Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, 4, 51-62.
4. Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., & Lin, C. T. (2017). A review of clustering techniques and developments. *Neurocomputing*, 267, 664-681.
5. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3), 264-323.

6. Madhulatha, T. S. (2012). An overview on clustering methods. *arXiv preprint arXiv:1205.1117*.
7. Petrelli, M. (2021). *Introduction to Python in Earth Science Data Analysis: From Descriptive Statistics to Machine Learning*. Springer Nature.
8. Masek, J. G., Wulder, M. A., Markham, B., McCorkel, J., Crawford, C. J., Storey, J., & Jenstrom, D. T. (2020). Landsat 9: Empowering open science and applications through continuity. *Remote Sensing of Environment*, 248, 111968.
9. Markham, B. L., Jenstrom, D., Masek, J. G., Dabney, P., Pedelty, J. A., Barsi, J. A., & Montanaro, M. (2016, September). Landsat 9: Status and plans. In *Earth Observing Systems XXI* (Vol. 9972, pp. 127-132). SPIE.
10. Huang, T. S., Schreiber, W. F., & Tretiak, O. J. (1971). Image processing. *Proceedings of the IEEE*, 59(11), 1586-1609.
11. Pandey, R., Naik, S., & Marfatia, R. (2013). Image processing and machine learning for automated fruit grading system: A technical review. *International Journal of Computer Applications*, 81(16), 29-39.
12. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering*, 9(03), 90-95.
13. Harris, C.R., Millman, K.J., van der Walt, S.J. et al. *Array programming with NumPy*. *Nature* 585, 357–362 (2020).
14. Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, *JMLR* 12, pp. 2825-2830, 2011.
15. Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., ... & Yu, T. (2014). scikit-image: image processing in Python. *PeerJ*, 2, e453.
16. Javed, A., Cheng, Q., Peng, H., Altan, O., Li, Y., Ara, I., ... & Saleem, N. (2021). Review of Spectral Indices for Urban Remote Sensing. *Photogrammetric Engineering & Remote Sensing*, 87(7), 513-524.
17. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202.
18. Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11), 559-572.

هشتمین همایش ملی انجمن رسوب شناسی ایران



8th Symposium of Sedimentological
Society of Iran / 31 Jan - 1 Feb 2024
University of Hormozgan

۱۱ و ۱۲ بهمن ماه ۱۴۰۲
دانشگاه هرمزگان

19. Daffertshofer, A., Lamoth, C. J., Meijer, O. G., & Beek, P. J. (2004). PCA in studying coordination and variability: a tutorial. *Clinical biomechanics*, 19(4), 415-428.
20. Kuleshov, V. (2013, May). Fast algorithms for sparse principal component analysis based on Rayleigh quotient iteration. In *International Conference on Machine Learning* (pp. 1418-1425). PMLR.
21. Parlett, B. N. (1974). The Rayleigh quotient iteration and some generalizations for nonnormal matrices. *Mathematics of Computation*, 28(127), 679-693.
22. Panda, S. K., Nag, S., & Jana, P. K. (2014, December). A smoothing based task scheduling algorithm for heterogeneous multi-cloud environment. In *2014 International Conference on Parallel, Distributed and Grid Computing* (pp. 62-67). IEEE.
23. Al Shalabi, L., Shaaban, Z., & Kasasbeh, B. (2006). Data mining: A preprocessing engine. *Journal of Computer Science*, 2(9), 735-739.
24. Patro, S. G. O. P. A. L., & Sahu, K. K. (2015). Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*.