ELM472 – Makine Öğrenmesinin Temelleri

Ödev 3

K-Means ile Salinas Haritası Üzerinde Sınıflandırma

Son teslim tarihi: 19.11.2022 – 23:59

Alican Bayındır
a.bayindir2020@gtu.edu.tr
Elektronik Mühendisliği Bölümü, GTÜ, Kocaeli, Türkiye

I. GİRİŞ

K-Means algoritması bir unsupervised learning(gözetimsiz öğrenme) ve kümeleme algoritmasıdır. 16 milyona kadar renk ve piksel başına 24 bit içerebilen bir resmimiz olduğunu düşünelim. Yalnızca 256 renk görüntüleyebilen 8 bit renkli bir ekranımız olduğunu varsayalım. 16 milyon renk arasından, sadece bu 256 rengi kullanarak orijinaline olabildiğince benzer bir resim üretecek 256 rengi seçmek istiyoruz. Bu noktada yüksek çözünürlükten düşük çözünürlüğe doğru eşleme yapıyoruz. Amaç, genel durumda sürekli bir uzaydan ayrık bir uzaya eşlemektir; bu vektör prosedürü, vektör niceleme olarak bilinir.

Elbette, eşit olarak niceleme yapabiliriz, ancak görüntüde bulunmayan renkler için girişler ekleyerek veya görüntüde yaygın olarak kullanılan renkler için daha fazla giriş eklemeyi atlayarak renk haritasını boşa harcamış oluruz. Örneğin, görüntü bir deniz manzarasıysa, hiç kırmızı görmeyi beklemeyebiliriz. Bu nedenle, renk haritası girişlerinin dağılımı, çok sayıda girişi yüksek yoğunluklu bölgelere yerleştirerek ve veri olmayan bölgeleri göz ardı ederek orijinal yoğunluğa mümkün olduğunca yakın olmalıdır. Bu bölgeler, rastgele seçilen bölgelerle aradaki mesafenin hesaplanması yolu ile kaç adet sınıf varsa o kadar sınıfa ayrılmasıyla bulunur.

II. UYGULAMA

K-means kümeleme, küme içi kareler toplamını azaltmak için bir dizi gözlemi (x1, x2,..., xn) k (n) S = S1, S2,..., Sk kümelerine bölmeyi amaçlar. (WCSS). Verilen kümedeki her gözlem, d-boyutlu bir gerçek vektördür. Resmi olarak, amaç şunları keşfetmektir:

$$argmin \sum_{i=1}^{k} \sum_{x \in Si} ||x - x_i||^2 = argmin \sum_{i=1}^{k} |S_i| Var S_i$$

Burada Si'nin ortalama puan değeri i'dir. Sonuç olarak, aynı küme içindeki noktaların ikili kare sapmaları en aza indirilir.

Aşağıdaki aşamalar, K-Means algoritmasının nasıl çalıştığını göstermektedir:

$$argmin \sum_{i=1}^{k} \frac{1}{|S_i|} \sum_{x,y \in Si} ||x - y||^2$$

Özdeşlik, aşağıdaki eşdeğerliğin çıkarılmasına izin verir.

$$|S_i| \sum_{x \in Si} ||x - u_i| \quad |^2 = \sum_{x \,!=y \in Si} ||x - y||^2$$

Adım 1: Küme sayısını belirlemek için K'yi seçin.

Adım 2: Rastgele K konum veya merkez noktası seçin. (Sağlanan veri kümesi olmayabilir.)

Adım 3: Her bir veri noktasını, önceden belirlenmiş olan K kümesini oluşturacak olan en yakın merkezine atayın.

Adım 4: Varyansı belirleyin ve her bir kümenin ağırlık merkezini yeniden konumlandırın.

Adım 5: Üçüncü adımı tekrarlayarak her veri noktasını her kümenin yeni merkezine yeniden atayın.

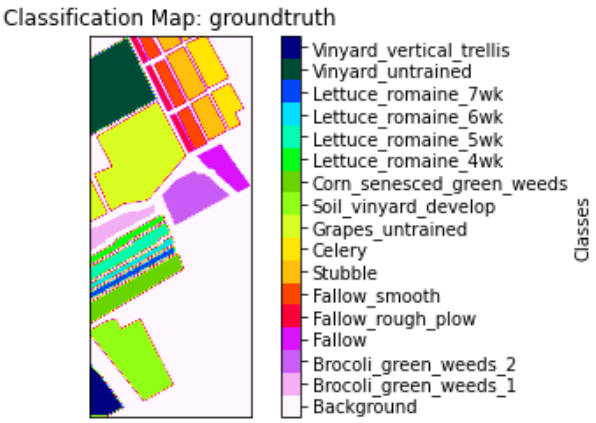
Adım 6: Yeniden atama varsa 4. adıma gidin; aksi takdirde bitişe gidin.

Adım 7: Bitmiş model.

A. Python kodunun yazılması

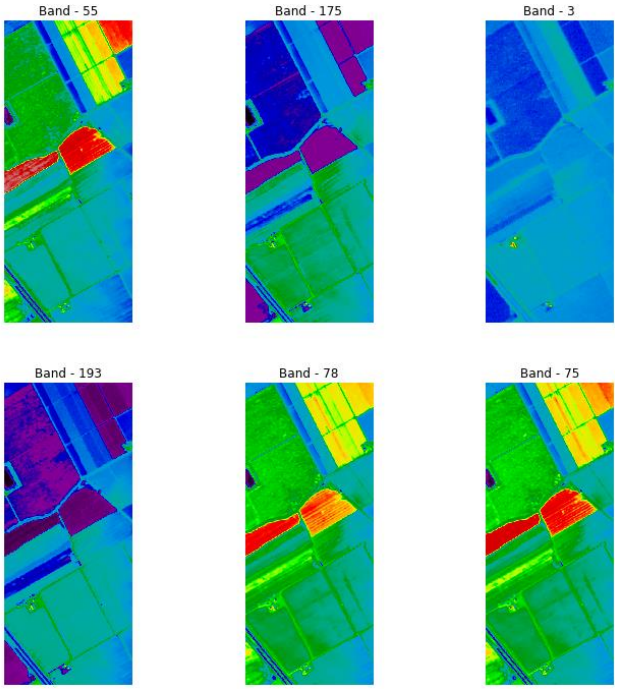
Gerekli araştırma ve konu incelenmesi yapıldıktan sonra Python kodu yazılmış ve ekte iletilmiştir. Formülün koda işlenmesi sırasında ilk olarak izin verilen kütüphaneler ve scikitlearn kütüphanesi çıkan sonucu kontrol amaçlı içeri aktarılmıştır.

ELM 472 – Makine Öğrenmesi Temelleri
Ödev 3
Salinas Haritası – K Means
- Alican Bayındır -



Şekil 1 Verilen Ground Truth dosyasının içeriği.

Daha sonra içe aktarılan veri dosyası boyutu 512, 217, 204 olması sebebiyle 204 katmandan rastgele 6 tanesi neler olduğunu görmek için çizdirilmiştir.



Şekil 2 Rastgele çizdirilen 204 katman içerisindeki 6 katman.

Dosya içerisindeki tüm bantların özellikleri describe metodu ile aşağıdaki şekilde gösterilmiştir.

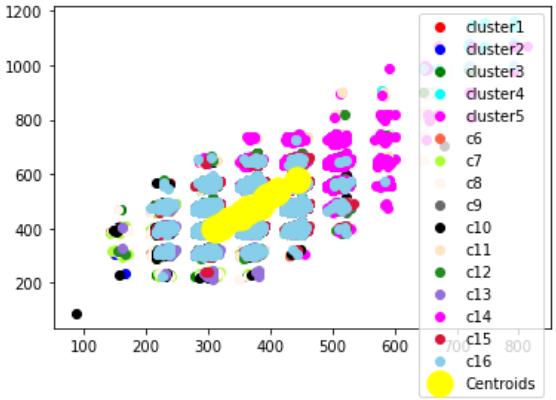
	0	1	2	3	4
count	111104.000000	111104.000000	111104.000000	111104.000000	111104.000000
mean	372.181929	480.388321	693.751287	1156.562923	1462.129950
std	62.453906	80.614130	103.217542	170.676450	227.876068
min	87.000000	86.000000	90.000000	86.000000	91.000000
25%	308.000000	404.000000	604.000000	1005.000000	1260.000000
50%	372.000000	482.000000	721.000000	1174.000000	1496.000000
75%	435.000000	558.000000	745.000000	1267.000000	1617.000000
max	814.000000	1165.000000	1920.000000	3865.000000	5153.000000

8 rows × 204 columns

Şekil 3 Her bir katmanın bazı özellikleri

III. SONUÇ

Sınıflandırma işleminden sonra elde edilen görüntü.



Şekil 4 Sınıflandırma sonrası elde edilen verilerin plot edilmesi sonucu oluşan görüntü.

Yukarıdaki görüntüde verilerin başarıyla sınıflandırıldığı düşünülmektedir. Vakit yetmediğinden dolayı % kaç başarı ile tespit edildiğini ve görüntünün son halini yazdırma kısmı başarılı bir şekilde sonuçlanmadı. Centroidler scikit ve kendi hesaplarımın neticesine bakıldığında yakın centroidler elde edilmiştir.

IV. KAYNAKÇA

[1] Ahmet Güneş, Dr. Öğr. Üyesi, ELM 472 - Makine Öğrenmesi Temelleri dersi, Gebze Teknik Üniversitesi.

[2] E. Alpaydin, Introduction to Machine Learning, 3. bs. Cambridge, MA, USA: MIT Press, 2014.

[3]https://scikitlearn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[4] Andrew NG, Dr. , CS 229 - Makine öğrenmesi, Harvard Üniversitesi

[5] https://github.com/meghshukla/LEt-SEN (a part in code)