ELM472 Makine Öğrenmesinin Temelleri
K-Means Öbekleme Ödev-3

Selimhan Aygün
s.aygun2019@gtu.edu.tr
Elektronik Mühendisliği Bölümü, GTÜ, Kocaeli, Türkiye

ÖZET

Bu çalışmada salinas verisi k-means öbekleme yöntemi kullanılarak öbeklenmiştir.

I. GİRİŞ

Çalışmada, k-means kümeleme algoritmasının temel prensiplerinin ve araştırılması amaçlanmaktadır. K-means, basit ve yaygın olarak kullanılan bir kümeleme algoritmasıdır. K-means algoritması, veri noktalarını k belirli sayıda küme veya grup içerisine yerleştirerek birbirine benzer olan noktaları aynı kümeye, farklı olanları ise farklı kümeler arasına dağıtmayı hedefler. Raporun devamında, k-means algoritmasının çalışma prensipleri, avantajları ve sınırlamaları detaylı olarak incelenecektir.

II. TEORİ VE YÖNTEM

Verilen veriler okunduktan sonra, verilerin 3 boyutlu olarak verildiği fark edildi. Verilerin daha kolay öbeklenmesi adına veriler 2 boyutlu olacak şekilde yeniden şekillendirildi.

Başlangıçta, k adet rastgele merkez nokta seçilir. Bu merkez noktalar, oluşturulacak küme sayısına bağlı olarak veri noktalarının bir araya gelmesini sağlayacak şekilde seçilmelidir. Her veri noktası, en yakın merkez noktaya atanır ve bu atandığı merkez noktasının bulunduğu kümeye işaretlenir. Her kümenin varyansı hesaplanır ve merkez noktası yeni ağırlık merkezine yerleştirilir. İşlem tekrarlanır, veri noktaları tekrar en yakın veri kümesine atanır. Atamalar değişmez hale geldiğinde, algoritma sonlanır ve kümeler elde edilmiş olur.

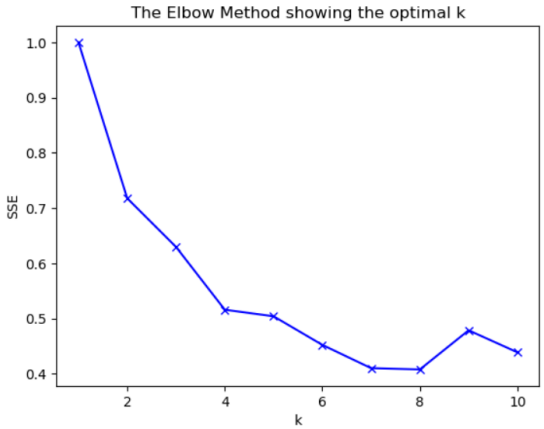
Merkezler ve veri noktaları arasındaki mesafe, k-means küme içi varyansı en aza indirgediğinden ve öklid uzaklığı metodu ile aynı olduğundan (2.1) numaralı denklemde de görünen mklid uzaklığı metodu kullanıldı.

$$d(p,q) = d(q,p)$$

$$= \sqrt{((q1-p1)^2 + (q2-p2)^2 + \cdots + (qn-pn)^2)}$$

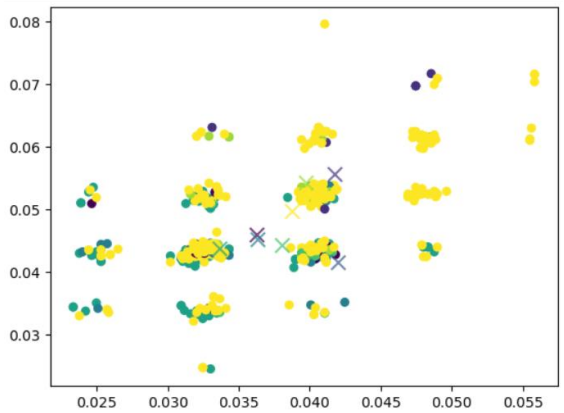
$$\sqrt{\sum_{i=1}^{n}(qi-pi)^2} \tag{2.1}$$

K-means algoritması büyük verilerdeki etkili sonucu, basit ve anlaşılır olması açısından avantajlara sahiptir. Ancak, k-means kümeleme algoritmasının bazı sınırlamaları da bulunmaktadır. İlk olarak, küme sayısı (k) önceden belirlenmelidir ve bu değer doğru şekilde seçilmelidir; aksi takdirde, yanlış kümeler elde edilebilir. Bunu gerçekleştirmek için çalışmada "dirsek metodu" (elbow method) diye adlandırılan yöntem kullanıldı ve optimum k değeri bulunmaya çalışıldı. Şekil 1'deki elbow metodu veride çok düzgün çalışmadı ve böyle bir sonuç verdi.



Şekil 1. Elbow metodu grafiği

k değerinin 8 olduğu yerde en az hata bulunduğundan ötürü k=8 olarak verinin öbeklenmesine devam edildi. K= 8'e göre öbeklenen veri Şekil 2'de verilmiştir.

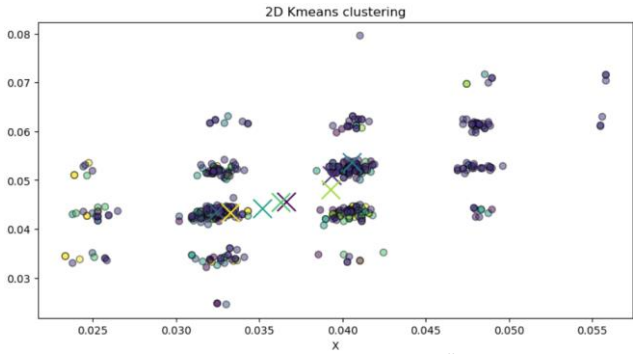


Şekil 2. Öbeklenmiş Veri

```
Clustering Error: 0.44925578470520267
```

Şekil 3. Öbekleme Hatası

Ayrıca, algoritma veri noktalarını yalnızca küme merkezlerine olan uzaklıklara dayanarak gruplandırır, bu da bazen yanlış atamalara veya düşük performansa neden olabilir.



Şekil 4. Sklearn Kütüphanesi Kullanılarak Öbeklenmiş Veri

Aynı veriler sklearn kütüphanesi ile öbeklendiğinde k=8 Şekil 4'teki gibi bir sonuç elde edilmiştir.

III. ANALİZ VE YORUM

Veriler yeniden şekillendirildikten sonra elde edilen verilerin öbekleme kodu doğru olsa bile tam düzgün olarak öbeklenmediği gözlemlenmiştir. Ayrıca, make_blob gibi başka verilerle test edildiğinde kodun düzgün olarak çalıştığı da gözlemlenmiştir. Verilerin yeniden şekillendirilmeden direkt verildiği şekilde 3 boyutlu olarak öbeklenmesi ne yazık ki başarılamamıştır. Verilerin sklearn kütüphanesi kullanılarak öbeklenmesinin de benzer oldugu görülmüştür.

Bu çalışmayla k-means öbekleme yöntemi araştırılmış ve uygulanmıştır.

KAYNAKÇA

[1] https://www.youtube.com/watch?v=5w5iUbTlpMQ
[2] https://github.com/llSourcell/k_means_clustering/blob/master/kmeans.py.ipynb
[3] https://medium.com/@sk.shravan00/k-means-for-3-variables-260d20849730
[4] https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/