**Kümelemede Normalleştirmenin Etkisi**

**Alper Şen1,\*, Türkay Gökgöz1**

*1Yıldız Teknik Üniversitesi, İnşaat Fakültesi, Harita Mühendisliği Bölümü, 34220, İstanbul.*

*Özet*

*Bu yazıda, hidrografik ağın k-ortalamalar yöntemiyle kümelenmesinde normalleştirme tekniklerinin etkisi incelenmiştir. Hidrografik nesnelerden türetilen geometrik ve topolojik öznitelikler, verinin dağılımını değiştirmeyen minimum-maksimum ve verinin dağılımını normale yaklaştıran lojistik teknikleri ile normalleştirilmiştir. Orijinal ve normalleştirilmiş veriler ile kümeleme sonuçları karşılaştırılmıştır.*

Anahtar Sözcükler

Veri Dönüşümü, Normalleştirme, Standartlaştırma, Kümeleme, K-Ortalamalar

**1. Giriş**

Kümeleme, verilerin kendi aralarındaki benzerliklerine göre gruplandırılması işlemidir (Han ve Kamber, 2001). Yaygın olarak kullanılan kümeleme yöntemlerinden biri olan K-ortalamalar yöntemi, *k* sayıda küme için kümelerin ortalamalarına göre nesneleri gruplayarak toplam uzaklığı minimize etmeyi amaçlamaktadır. Kümeleme analizi, matematik, istatistik, bilgisayar bilimi, makine öğrenimi, yapay zeka, veri madenciliği, görüntü işleme, kartografya ve coğrafi bilgi sistemleri gibi alanlarda sıklıkla kullanılmaktadır.

Normalleştirme, veri madenciliğinin önemli bir konusu olan kümelemenin temel ön işlem adımlarından biridir. Normalleştirme tekniklerinden hangisinin kullanılacağı, özniteliğin istatistiksel tanımı ve uygulamanın amacına göre değişiklik göstermektedir. Özniteliklerin ortalama ve varyansları birbirinden önemli ölçüde farklı olduğu takdirde büyük ortalama ve varyansa sahip özniteliklerin diğerleri üzerindeki etkisi daha fazla olur ve onların rollerini önemli ölçüde azaltır. Ayrıca özniteliklerin sahip olduğu çok büyük ve çok küçük değerler de analizin sağlıklı biçimde gerçekleştirilmesini etkiler. Bu nedenle bir dönüşüm tekniği uygulayarak söz konusu özniteliklerin normalleştirilmesi veya standartlaştırılması uygun bir yol olur (Özkan, 2008). K-ortalamalar kümeleme yönteminde kullanılan Öklid uzaklık kriteri, öznitelik büyüklükleri arasındaki uyumsuzluğa son derece duyarlıdır (Patel ve Mehta, 2011). Bu bağlamda, uygun normalleştirme tekniğinin uygulanması, kümeleme kalitesini artıracaktır (Karthikeyani vd., 2009).

Minimum-Maksimum (M-M) normalleştirme tekniğinde veriler en büyük ve en küçük değerlere göre 0-1 aralığında yeni değerlere dönüştürülür. İstatistikte sıkça kullanılan normalleştirme tekniklerinden bir diğeri olan Z-skor standartlaştırmasında veriler, aritmetik ortalaması sıfır ve varyansı bir olacak şekilde yeni değerlere dönüştürülür. M-M normalleştirme ve Z-skor standartlaştırma doğrusal teknikler olup, verinin dağılımını değiştirmemektedir. Buna karşın, logaritmik normalleştirme, lojistik normalleştirme ve histogram eşitleme doğrusal olmayan tekniklerdir ve verinin dağılımını değiştirmektedir. Değişkenlerin değeri üssel olarak dağıldığı durumlarda logaritmik normalleştirme uygun olacaktır. Bu tekniğin dezavantajı değeri sıfır veya negatif olan veriye uygulanamayışıdır. Bu dezavantaj bazı çalışmalarda veri pozitif hale dönüştürülerek giderilmektedir. Lojistik normalleştirme, lojistik fonksiyonları kullanarak veriyi 0-1 aralığına dönüştürmektedir. Bu dönüşüm ortalama etrafında az ya da çok doğrusallık gösterirken, uç değerlerde yumuşak bir doğrusal olmama özelliği taşımaktadır. Histogram eşitleme tekniğinde ise veriler büyüklüklerine göre sıralanır ve her bir değer sıra numarası ile değiştirildikten sonra 0-1 aralığına dönüştürülür. Sürekli ve kesikli değişkenler için kullanışlıdır [1].

Karthikeyani vd. (2009) tarafından üç farklı normalleştirme tekniği (M-M, Z-skor ve ondalık ölçeklendirme (decimal scaling)), K-ortalamalar kümeleme yönteminde karşılaştırılmıştır. Wu vd. (2009), K-ortalamalar yöntemi ile yapılan kümeleme performansının değerlendirilmesinde normalleştirmenin önemine dikkat çekmiştir. Patel ve Mehta (2011), K-ortalamalar kümeleme yönteminin performansını, aykırı değerlerin tespiti ve normalleştirme açısından değerlendirmiştir. Souto vd. (2008), üç farklı normalleştirme tekniğini (M-M, Z-skor ve rank), beş farklı kümeleme yönteminde karşılaştırmıştır. Bu çalışmalarda, kümelemenin normalleştirme tekniklerine duyarlı olduğu belirlenmiştir.

Bu çalışmada, USGS hidrografi veri seti kullanılmıştır. Hidrografik öznitelikler, geometrik, topolojik ve semantik olmak üzere üç grupta toplanmaktadır. K-ortalama algoritması sadece sürekli veriler ile çalışabilmekte, kategorik veriler ile çalışamamaktadır (Huang, 1998). Bu yüzden sürekli öznitelikler kullanılmıştır. Kullanılan geometrik öznitelikler, uzunluk ve kıvrımlılık; topolojik öznitelikler ise bağlılık, arasında olma ve yakınlık bakımından merkezlik değerleridir. Hidrografik nesnelerden türetilen geometrik ve topolojik öznitelikler, verinin dağılımını değiştirmeyen M-M tekniği ve verinin dağılımını normale yaklaştıran lojistik tekniği ile normalleştirilmiştir. Orijinal ve normalleştirilmiş öznitelikler istatistiksel olarak incelenmiştir. Normalleştirme tekniklerinin kümelemeye etkisi niceliksel ve görsel olarak tespit edilmiştir.

**2. K-Ortalamalar Kümeleme Yöntemi**

Hiyerarşik olmayan kümeleme yöntemleri arasında önem taşır ve yaygın biçimde kullanılır. K-ortalama algoritması, büyük veri setlerini kümelemedeki etkinliği ile bilinmektedir. Bu yöntemde, belli sayıdaki küme için toplam uzaklık değerini minimize etmek amaçlanır.

K-ortalamalar algoritmasında öncelikle *k* küme sayısının belirlenmesi gerekir. Söz konusu *k* değeri belirlendikten sonra her bir kümeye gözlem değerleri rastgele olarak atanır ve böylece *Cs* (*s=1,2,…,k*) kümeleri belirlenmiş olur. Ardından aşağıdaki işlemler gerçekleştirilir.

1. Her bir kümenin merkezi *Ms* (*s=1,2,…,k*) belirlenir.
2. Her bir küme içinde her bir gözlem değerinin merkeze olan uzaklıklarının toplamı *es* (*s=1,2,…,k*) hesaplanır. Bütün kümeler için toplam uzaklık değeri *E* bulunur.
3. *Ms* (*s=1,2,…,k*) merkez değerleri ile gözlem değerleri arasındaki uzaklıklar hesaplanır. Bir gözlem değeri hangi merkeze yakın ise, o merkeze sahip küme içine dahil edilir.
4. Yukarıdaki a, b ve c adımları, kümelerde herhangi bir değişiklik olmayıncaya kadar tekrar edilir.

Algoritmada (*s=1,2,…,k*) olmak üzere *Cs* kümesinin *Ms* merkezi, (1) eşitliği ile hesaplanır.

 (1)

Burada *xis* değeri *Cs* kümesine ait *i*. gözlemdir. *Cs* kümesi *ns* sayıda gözlem içermektedir. *Cs* kümesi için toplam uzaklık *es* her bir *g*özlem *xis* ile onun merkezi *Ms* arasındaki Öklid uzaklıkları toplamıdır ve (2) eşitliği ile hesaplanır.

 (2)

Bütün kümeler için toplam uzaklık *E* ise (3) eşitliği ile hesaplanır.

 (3)

**3. Uygulama**

* 1. **Veri Kaynağı ve Çalışma Alanı**

Bu çalışmada, United States Environmental Protection Agency (USEPA) ve United States Geological Survey (USGS) tarafından birçok veri kaynağı ve web harita servisinden derlenen hidrografik veri setlerinden (National Hydrography Dataset) biri olan 1:24,000 ölçekli “NHDH1029” isimli veri setinin içerdiği “Pomme de Terre” adlı havzanın “Crane” adlı alt havzasına ait veriler kullanılmıştır. Pomme De Terre, Missouri eyaletinde, iç dağlık bölgesi olan Ozark Platosu’nda bulunmaktadır.

Crane alt havzasındaki akarsu nesnelerine ait öznitelikler türetilmeden önce, akarsulara ait segmentler, veri tabanında bulunan akarsu tipi (daimi ve kuru dere) ve akarsu düzeyi bilgilerine ((Horton (1945) ve Strahler (1957)) göre birleştirilerek, ana akarsular ve kolları oluşturulmuştur. Crane alt havzasında, 122 adet akarsu bulunmaktadır.

* 1. **Öznitelikler**

**3.2.1. Geometrik Öznitelikler**

**Uzunluk:** Her bir akarsu nesnesinin uzunluğudur.

**Kıvrımlılık:** Her bir akarsuyun başlangıç ve bitiş noktalarının koordinatları belirlenerek, iki nokta arasındaki uzunluk (Öklid uzunluğu) hesaplanmış ve Öklid uzunluğunun akarsuyun uzunluğuna oranı ile kıvrımlılık değerleri bulunmuştur. Şekil 1.’de görünen d1/d2 oranı, kıvrımlılık değerini vermektedir.

kıvrımlılık.tif

*Şekil 1: Kıvrımlılık özniteliği*

**3.2.2. Topolojik Öznitelikler**

Günümüzde, ağ yapısı oluşturan nesnelerin (örneğin; akarsu ağı, yol ağı, vb.) topolojik olarak modellenmesi çizge (graph) üzerinden sağlanmaktadır. Topolojik model, çizgenin düğüm noktaları ve kenarları ile oluşturulmaktadır (Gülgen, 2009).

Merkezlik (centrality) kavramı ilk defa 1948 yılında Bavelas tarafından sosyal ağlarda insan iletişimi konulu bir uygulama yapmak amacıyla ortaya atılmıştır. Bavelas, merkezlik ve grup içi iletişim arasında bir ilişki olduğu hipotezi ile yola çıkmıştır. Merkezlik kavramı, sosyal ağların önemli bir yapısal özelliğidir. Bir ağdaki noktaların merkezlik değerini tespit eden üç önemli ölçüt aşağıda verilmektedir (Freeman, 1978). Merkezlik değerleri çizgeden Pajek yazılımı (Batagelj and Mrvar, 2008) ile elde edilmiştir.

**Bağlılık bakımından merkezlik (Degree centrality):** Bir akarsuya bağlı akarsuların sayısı veya çizgede bu akarsuya karşılık gelen düğüm ile bağlantılı düğüm sayısıdır. Verilen bir *pk* düğümünün bağlılık bakımından merkezlik değeri, (4) eşitliği ile hesaplanmaktadır (Freeman, 1978; Jiang ve Harrie, 2004; Gülgen, 2009).

 (4)

*pi* ve *pk* bir kenar ile bağlı ise, ; bağlı değil ise 0 değerini alır.

**Yakınlık bakımından merkezlik (Closeness centrality):** Çizge üzerinde iki düğüm noktası arasındaki en kısa güzergâhta (shortest linking) bulunan kenar sayısına (number of edges) bağlı olarak değişen bir ölçüttür. Çizgede herhangi iki düğüm noktası olan *pi* ve *pk* noktaları arasındaki en kısa güzergâhta bulunan kenar sayısı *d(pi, pk)* ile gösterildiğinde *pk* noktasının yakınlık bakımından merkezlik değeri, (5) eşitliği ile hesaplanmaktadır (Freeman, 1978; Jiang ve Harrie, 2004; Gülgen, 2009).

 (5)

**Arasında olma bakımından merkezlik (Betweenness centrality):** Bir akarsudan diğer akarsulara olan en kısa güzergâhlarda, arada uğranılması zorunlu olan akarsu için arasında olma bakımından merkezlik ölçütü kullanılabilir. Çizgede herhangi iki düğüm noktası olan *pi* ve *pj* noktaları arasındaki *pk* noktasından geçen en kısa güzergahların sayısı hesaplanarak bulunan, *pk* noktasının arasında olma bakımından merkezlik değeri, (6) eşitliği ile hesaplanmaktadır (Freeman, 1978; Jiang ve Harrie, 2004; Gülgen, 2009).

 (6)

*gij*: *i* ve *j* düğümleri arasındaki en kısa güzergah sayısı

*gij(pk)*: Bir *pk* noktasından geçilmesi zorunlu olması durumunda, *i* ve *j* düğümleri arasındaki en kısa güzergah sayısı

* 1. **Normalleştirme ve Kümeleme**

M-M normalleştirme tekniği ile veriler en büyük ve en küçük değerlere göre 0-1 aralığında yeni değerlere dönüştürülmüştür. M-M normalleştirmesinde (7) eşitliği kullanılmıştır.

 (7)

Burada , normalleştirilen değerleri; *Xi,* öznitelik değerlerini; *Xmin*, en küçük öznitelik değerini ve *Xmaks*, en büyük öznitelik değerini ifade etmektedir.

Lojistik normalleştirme işleminde öncelikle Z-skor standartlaştırması ile öznitelikler standartlaştırılmıştır (Eşitlik 8). Daha sonra, (9) eşitliğiyle lojistik fonksiyon uygulanmıştır.

 (8)

 (9)

Burada , örneklemin aritmetik ortalaması ve *s,* standart sapmadır.

Kümeleme, Matlab R2009 fonksiyonları yardımıyla, K-ortalamalar kümeleme yöntemine göre, hem orijinal hem de normalleştirilmiş (M-M ve lojistik teknikleriyle) veriler ile gerçekleştirilmiştir. Kümeleme sonuçlarının karşılaştırılabilmesi için her üç kümelemede de aynı küme sayısı (*k=5*) seçilmiştir.

* 1. **Bulgular**

Orijinal ve normalleştirilmiş özniteliklerin dağılımlarını incelemek amacıyla Şekil 2’de belirtilen histogram grafikleri elde edilmiştir. Histogram grafiklerinde, lojistik normalleştirmede dağılımının normal dağılıma yaklaştığı görünmektedir.

dağılımlar.tif

*Şekil 2: Lojistik normalleştirmenin veri dağılımına etkisi*

Özniteliklerin normal dağılımlı olup olmadığını belirlemek için Kolmogorov-Smirnov (K-S) uygunluk tek örnek testi, test büyüklüğü (D) ve Smirnov’un kritik tablo değerleri (dα) ile yapılmıştır (anlamlılık düzeyi α=0.05). Sıfır hipotezi (gözlemler normal dağılımlıdır) reddedildiğinden (dα<Dorj ve dα<Dloj), öznitelikler normal dağılımlı değildir. Bununla birlikte, öznitelikler lojistik normalleştirme ile dönüştürüldükten sonra –orijinal verilere göre- normal dağılıma yaklaşmıştır (Tablo 1).

*Tablo 1: Orijinal ve lojistik teknik ile normalleştirilmiş verilerle K-S uygunluk tek örnek testi*

|  |  |  |  |
| --- | --- | --- | --- |
| **Öznitelik** | **Dorj** | **Dloj** | **d0.05** |
| Uzunluk | 0.2621 | 0.1921 | 0.1215 |
| Kıvrımlılık | 0.1508 | 0.143 |
| Arasında | 0.3957 | 0.338 |
| Yakınlık | 0.1707 | 0.1676 |
| Bağlılık | 0.3766 | 0.3518 |

Orijinal ve iki farklı teknikle normalleştirilmiş öznitelikler kullanılarak yapılan K-ortalama kümelemesi farklı sonuçlar ortaya koymuştur. K-ortalamalar yönteminde *k=5* için elde edilen toplam uzaklık değerleri (*E*), orijinal veriler ile 76.84, lojistik tekniğine göre normalleştirilmiş veriler ile 4.22 ve M-M tekniğine göre normalleştirilmiş veriler ile 2.93 olarak bulunmuştur. Orijinal özniteliklerin sahip olduğu çok büyük ve çok küçük değerler, toplam uzaklık değerinin diğerlerine göre daha fazla olmasına yol açmıştır. Şekil 3’de üç farklı kümeleme sonucu görünmektedir.

kumeleme.tif

*Şekil 2: Kümelemede normalleştirmenin etkisi*

Sonuç olarak, orijinal veriler ile kümelemede, uzunluk özniteliğinin baskın sayısal değerleri kümelemeyi etkilemiştir. Beş kümenin ikisi, en uzun nesnelerden oluşan birer elemana sahiptir (kırmızı ve mor).

M-M tekniğine göre normalleştirilmiş veriler ile kümelemede, genellikle tüm öznitelikler göz önüne alınmıştır. Mavi ile gösterilen küme daha çok kıvrımlılık özelliği gösterirken, mor ile gösterilen küme daha çok uzunluk, bağlılık bakımından merkezlik ve arasında olma bakımından merkezlik özelliklerini göstermektedir. Sarı ile gösterilen küme daha çok uzun olan nesneleri içermektedir. Yeşil ile gösterilen küme ise daha çok yakınlık bakımından merkezlik özelliğini göstermektedir.

Lojistik tekniğine göre normalleştirilmiş veriler ile kümeleme sonuçlarında, mor ile gösterilen küme daha çok en uzun ve arasında olma bakımından merkezlik özelliği gösteren nesneleri içermektedir. Sarı ve yeşil ile gösterilen kümeler daha çok yakınlık bakımından merkezlik özelliği göstermektedir.

1. **Sonuç ve Öneriler**

Normalleştirme, veri madenciliğinin önemli konularından biri olan kümelemenin temel ön işlem adımlarından biridir ve normalleştirme tekniklerinden hangisinin kullanılacağına, özniteliğin istatistiksel tanımı ve uygulamanın amacına göre karar verilmelidir. Örneğin, aykırı değerlerin ortalama değerlerin yer aldığı kümelere dahil olmaması, ayrı kümlerde yer alması istendiğinde, M-M normalleştirme tekniği tercih edilebilir. Lojistik tekniği ile normalleştirmede aykırı değerlerin ortalama değerlere yaklaşması ve dolayısıyla ortalama değerlerin yer aldığı kümelere dahil olması olasıdır.

**Kaynaklar**

Batagelj, V. ve Mrvar, A., (2008). Networks/Pajec: Program for large networks analysis. <http://pajek.imfm.si/doku.php>.

Freeman, L.C., (1978). *Centrality in Social Networks Conceptual Clarification*, Social Networks, (1), Elsevier, 215-239.

Gülgen, F. (2009). *Yerleşim içi yol ağı genelleştirmesinde yeni bir seçme/eleme yaklaşımı,* Doktora Tezi. Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Han, J. and Kamber, M., (2001). *Data mining: Concepts and Techniques*, Morgan Kaufmann Publishers, USA.

Horton, R.E., (1945). *Erosional development of streams and their drainage basins; hydrophysical approach to quantitative morphology*, Bulletin of the Geological Society of America, 56(3), 275-370.

Huang, Z., (1998). *Extentions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values*, Data Mining and Knowledge Discovery, 2, 283-304.

Jiang, B. ve Harrie, L., (2004). *Selection of streets from a network using self-organizing maps*, Transactions in GIS, 8(3), 35-350.

Karthikeyani, V. ve Thangavel, K., (2009). Impact of nornalization in distributed K-means clustering, Int. Jour. of Sof Computing, 4(4), 168-172.

Strahler, A.N., (1952) *Dynamic basis of geomorphology*, Bulletin of Geological Society of America, 63(7), 923-938.

Özkan, Y., (2008). *Veri Madenciliği Yöntemleri*, Papatya Yayıncılık, İstanbul.

Patel V.R. ve Mehta R.G., (2011). Impact *of outlier removal and normalization approach in modified K-means clustering algorithm*, Int. Jou. of Computer Science Issues, 8(5), 331,336.

Souto, M.C.P., Araujo, D.A.S., Costa, I.G., Soures, R.G.F., Ludermir T.B. ve Schliep, A., (2008). *Comparative study on normalization procedures for cluster analysis of gene expression datasets*, Proc. IEEE Int. Joint Conferece Neural Networks, Hong Kong, 2792-2798.

Wu J., Xiong, H. ve Chen, J., (2009). Adapting the Right Measures for K-means Clustering. Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 877-885, Paris.

[1] http://www.cis.hut.fi/somtoolbox/