International Journal of Scientific and Technological Research www.iiste.org
ISSN 2422-8702 (Online), **DOI:** 10.7176/JSTR/7-01-05
Vol.7, No.1, 2021

Determination of Initial Centers in K-Means Clustering Method by NAMGY Algorithm

Meryem Goral Yildizli (Corresponding author)
Department of Biostatistics, Faculty of Medicine,
Cukurova University, Adana, Turkey
E-mail: meryem@cu.edu.tr

Zeliha Nazan Alparslan
Department of Biostatistics, Faculty of Medicine,
Cukurova University, Adana, Turkey
E-mail: nazan@cu.edu.tr

Abstract
Objective: With the development and widespread use of technology, the increasing volume of data in many areas has accelerated the digitization process. The gains obtained by processing and interpreting data stacks can make significant contributions to institutions and organizations in many managerial issues from production to decision-making processes. It has increased the use of data mining methods in different areas, which support the process of transforming digitalized large-scale data into information. One of the increasingly popular techniques in data mining is clustering, and the K-means algorithm is a non-hierarchical clustering method compatible with large amount of data. This method is widely used in the scientific studies, however the number of clusters and initial centers defined as parameters comes up a disadvantage for the algorithm, especially for those not familiar with the mathematical specificities. Initial centers those generated randomly by K-means usually make the clustering results reaching non-optimal. K-means algorithm is very sensitive in initial centers. More consistent results of K-means clustering can be achieved after computing more than one times. However, it is difficult to decide the computation limit, which can give the optimal result. An improvement of K-means algorithm with this respect will be a contribution on overcoming this disadvantage for scientific studies. In order to solve this problem; NAMGY (Neighborhood and Midpoint Gain Yield) algorithm has been developed, which includes methods that provide optimal selection of parameters according to the properties of objects. This article covers the application of the method of determining the initial centers in NAMGY algorithm.

Method: In order to analyze the accuracy of our proposed method, both the standard K-means and NAMGY algorithm were applied on the classified data set those Iris, Yeast and Segment-challenge. And also the performances of the algorithms in terms of the working principle were evaluated on the VitaminB12 data set obtained from the Cukurova University Balcalı Hospital Information Management System. Euclidean distances were calculated between objects and data sets were transformed into values in the range [0, 1] using normalization. Adjusted Rand index was used to evaluate the validity of clusterings.

Results: According to the examined results; the applications that reveal the effects of the initial centers on the analysis process of the algorithms have been carried out with different approaches such as the working principle of the algorithm, the effect of the initial centers on the clustering results, the evaluation of the clustering performance. It was again concluded that professional selection of parameters is requirement to increase the usability of a clustering algorithm and the reliability of clustering results. The NAMGY algorithm uses a systematic way to find initial centers which reduces the number of dataset scans and will produce better accuracy in smaller number of iteration. NAMGY algorithm has proved to be better than traditional K-means algorithm in terms of good quality results and analysis processes. According to the results generated; NAMGY provides a challenging algorithm for the disadvantage of the standard K-mean algorithm. However further research is required to verify the capability of this algorithm when applied to data sets with more complex objects.

Keywords: Clustering, K-means, Initial center, Data mining

DOI: 10.7176/JSTR/7-01-05

International Journal of Scientific and Technological Research www.iiste.org
ISSN 2422-8702 (Online), **DOI:** 10.7176/JSTR/7-01-05
Vol.7, No.1, 2021

K-ortalama Kümeleme Yönteminde Başlangıç Merkezlerinin NAMGY Algoritması ile Belirlenmesi

Özet

Amaç: Teknolojinin gelişmesi ve yaygın kullanılmasıyla birlikte birçok alanda hacmi giderek artan veriler dijitalleştirme sürecini hızlandırmıştır. Bu değişimle beraber veri yığınlarının işlenmesi ve yorumlanması sayesinde edilen kazanımlar üretimden karar verme süreçlerine kadar birçok alanda kurumlara ve kuruluşlara önemli katkılar sağlayabilir. Dijitalleşen büyük boyuttaki verilerin bilgiye dönüştürülme sürecine destek sağlayan veri madenciliği yöntemlerinin farklı alanlardaki kullanımı artmıştır. Veri madenciliğinde popülaritesi gittikçe artan tekniklerden biri kümeleme yöntemidir. Hiyerarşik olmayan kümeleme yöntemlerinden K-ortalama algoritması bilimsel çalışmalarda yaygın olarak kullanılmaktadır. Ancak K-ortalama yönteminin algoritmik parametre değerleri (küme sayısı, başlangıç merkezleri) ile farklı performans sonuçlarının oluşabilmesi algoritma için dezavantajdır. Rastgele seçilen farklı başlangıç merkezleriyle oluşturulan küme sonuçlarında tutarsızlıklar olabileceği gibi analiz sürecinde de algoritmanın uygulanma tekrar sayısını artırabilir. Bu durum bilimsel çalışmaların güvenirliğini azaltabilir ve büyük veri niteliğindeki veri setlerinin analiz süresini artırabilir. Problemin çözümüne yönelik geliştirilen algoritmalar, K-ortalama algoritmasının kullanımını artıracak, bilimsel çalışmalardan elde edilen sonuçların daha geçerli olmasına katkı sağlayacaktır. Parametre değerlerinin kullanıcıdan bağımsız belirlendiği K-ortalama tabanlı algoritma önerilmiştir. Standart K-ortalama algoritmasına veri setindeki nesnelerin özelliklerine göre uygun küme sayısını ve başlangıç merkezlerini belirleyen iki ayrı metot eklenerek NAMGY (Noktalar Arası Mesafe ve Gözlemlerin Yoğunluğu) isimli algoritma geliştirilmiştir. Bu makale NAMGY algoritmasının içerdiği metotlardan yalnızca başlangıç merkezlerinin belirlendiği yöntemin uygulamasını içermektedir.

Yöntem: Yeni Zelanda Waikato Üniversitesi tarafınca geliştirilen açık kaynak kodlu WEKA referans program olarak kullanılmıştır. NAMGY(K-ortalama) ve WEKA(K-ortalama) algoritmaları bilimsel çalışmalarda sıkça kullanılan İris, Yeast ve Segment-Challenge veri setleri üzerinde uygulanmıştır. Ayrıca iki algoritmanın çalışma prensibi açısından karşılaştırılmasında Çukurova Üniversitesi Balcalı Hastanesi Hastane Bilgi Yönetim Sisteminden alınan VitaminB12 veri seti üzerinde algoritmalar uygulanarak sonuçlar oluşturulmuştur. Veri setleri üzerinde normalizasyon işlemi yapılarak veri setlerindeki nesneler [0,1] aralığında değerlere dönüştürülmüştür. Nesneler arasındaki uzaklık ölçümlerinde Öklid uzaklık ölçütü kullanılmıştır. Çalışmada, dışsal indekslerden Düzeltilmiş Rand indeks küme geçerlilik ölçütü olarak kullanılmıştır.

Bulgular: Algoritmaların veri setleri üzerine uygulama sonuçları üç başlık altında sunulmuştur. **Algoritmanın çalışma prensibi;** NAMGY algoritmasının analiz sürecindeki işlemler açısından standart K-ortalama algoritmasıyla karşılaştırılması yapılmıştır. İki algoritmanın uygulama sonuçlarına göre çalışma prensibi açısından farlılıklar gözlenmiştir. NAMGY (K-ortalama) algoritmasında başlangıç merkezleri nesnelerin özellikleri dikkatte alınarak algoritmanın içerdiği yöntemle ve alan deneyimi gerektirmeden programın bir kez çalıştırılarak optimal küme sonuçları oluşturulmuştur. Oysa WEKA (K-ortalama) uygulanmasında uygun başlangıç merkezlerinin belirlenmesi, farklı seed değerleriyle 18 kez çalıştırılarak gözlemsel olarak optimal küme sonuçları oluşturulmuştur. Standart K-ortalama algoritmasında geçerli küme sonuçları oluşturmak için tekrarlı denemelerin yapılmasının gereksinim olduğu görülmüştür.
Başlangıç merkezlerinin küme sonuçlarına Etkisi; Farklı başlangıç merkezleri ile K-ortalama algoritmasının oluşturduğu değişken küme sonuçları gözlemsel olarak değerlendirilerek optimal kümeler belirlenmiştir. Algoritmanın küme sonuçları niteliksel olarak karşılaştırılarak, başlangıç merkez seçiminin küme sonuçlarına etkisi araştırılmıştır. Aynı veri setinin farklı SSE değerlerine göre oluşturulan kümelerin öne çıkardığı sonuçların birbiriyle tutarsız olduğu görülmüştür.
Kümeleme performanslarının değerlendirilmesi; NAMGY algoritması ve standart K-ortalama algoritmasının küme sonuçları küme geçerlilik indekslerine göre değerlendirilmiştir. NAMGY algoritmasında her üç veri seti için DRI değerleri daha yüksek olduğu, aynı zamanda daha düşük iterasyon ile daha düşük SEE değerleri bulunduğu görülmüştür. Üç performans kriterine göre NAMGY algoritmasının daha avantajlı olduğu söylenebilir.

Sonuc: NAMGY algoritmasının daha efektif analiz süreci ve kümeleme sonuçları oluşturduğu görülmüştür. Önerdiğimiz algoritmada parametre seçiminin profesyonel olarak yapılması algoritmanın kullanılabilirliğini ve kümeleme sonuçlarının güvenirliliğini artırması açısından önemlidir. Standart K-ortalama algoritmasının dezavantajı olan başlangıç merkezlerini belirleme problemine NAMGY bir çözüm seçeneğidir. NAMGY algoritmasının çeşitli kaynaklardan ve farklı formattaki çok boyutlu karmaşık nesnelerin oluşturduğu veri setleri üzerindeki uygulamaları ve performans araştırılması sonraki çalışmaların konusu olabilir.

Anahtar Kelimeler: Kümeleme, K-ortalama, Başlangıç Merkezler, Veri Madenciliği

1. Giriş

Teknolojinin gelişmesi ve yaygın kullanılmasıyla birçok alanda hacmi giderek artan veriler dijitalleştirme sürecini hızlandırmıştır. Bu değişimle beraber verilere işlerlik kazandırılması üretimden karar verme süreçlerine kadar birçok alanda kazanımlar sağladığı gibi sektörler için önemli güç haline gelmiştir. Verilerin bilgiye dönüştürülme sürecine destek sağlayan Veri madenciliği; geleneksel istatistiksel yöntemlerle beraber büyük veri niteliğindeki verileri işleyebilecek algoritmalarla entegre olmuş teknolojidir [1,10]. Dijital dönüşümle beraber veri madenciliği yöntemlerinin her alandaki kullanımı artmıştır. Veri madenciliğinde popülaritesi artan tekniklerden biri kümeleme yöntemidir. Kümeleme algoritmalarının temelini oluşturan K-ortalama algoritması, çeşitli veri setlerine uygulanabilir olması, farklı alanlara ilişkin bilimsel çalışmalarda tercih edilmesine katkı sağlamıştır. Ancak K-ortalama da algoritmik parametre değerleri (küme sayısı, başlangıç merkezleri) ile farklı performans sonuçlarının oluşabilmesi algoritmanın dezavantajıdır. Rastgele seçilen farklı başlangıç merkezleriyle oluşturulan küme sonuçlarında tutarsızlıklar olacağı gibi analiz sürecinde de algoritmanın uygulanma tekrar sayısını artırabilir. Bu durum bilimsel çalışmaların güvenirliğini azaltabilir ve büyük veri niteliğindeki veri setlerinin analizinde önemli zaman kaybıdır. Problemin çözümüne yönelik geliştirilen algoritmalar, K-ortalama algoritmasının kullanımını artıracak, bilimsel çalışmalardan elde edilen sonuçların daha geçerli olmasına katkı sağlayacaktır. Parametre değerlerinin kullanıcıdan bağımsız belirlendiği K-ortalama tabanlı algoritma önerilmiştir. Standart K-ortalama algoritmasına veri setindeki nesnelerin özelliklerine göre uygun küme sayısını ve başlangıç merkezlerini belirleyen iki ayrı metot eklenerek NAMGY (Noktalar Arası Mesafe ve Gözlemlerin Yoğunluğu) isimli algoritma geliştirilmiştir. Bu makale NAMGY algoritmasının içerdiği metotlardan yalnızca başlangıç merkezlerinin belirlendiği yöntemin uygulamasını içermektedir.

2. Literatür Taraması

K-ortalama yönteminde algoritmik parametrelerinin optimal belirlenmesini içeren bazı bilimsel çalışmalar aşağıda özetlenmiştir.

Dalhatu ve Sim [2] çalışmasında; K-ortalama algoritması için başlangıç merkezlerinin seçimine yönelik önerilen algoritmada, veri setindeki noktalar arasındaki en büyük ve en küçük uzaklıklara göre hesaplanan eşik değeri ve noktanın yoğunluk değeri dikkate alınmıştır. Önerilen algoritmanın, mevcut K-ortalama algoritmasına göre daha iyi performans gösterdiği belirtilmiştir.

Çolak ve arkadaşları [3] K-ortalamanın optimal küme sayısının seçiminde toplam hata kare kriteri kullanılmıştır. Matlab ortamında farklı veri setleri üzerinde oluşturulan uygulama sonuçlarına göre kümeleme performansının iyi olduğu belirtilmiştir.

Godara ve Sharma [4] çalışmasında; öncelikle başlangıç merkezleri en küçük örten ağaç (Minimal Spanning Tree (MST)) sınıflandırma yöntemi ile belirlenip, daha sonra K-ortalama algoritması seçilen merkezlerle uygulanarak kümeler oluşturulmuştur. Önerilen iki aşamalı yöntemin performansı klasik K-ortalama algoritmasına göre daha iyi olduğu vurgulanmıştır.

Kedar ve Sawant [5] çalışmasında; K-ortalama algoritmasında başlangıç merkezlerinin optimal seçimi için noktalar arasındaki uzaklık toplamlarının kullanılması önerilen algoritmanın tekrar (iterasyon) sayısını azalttığını ve sonuçların daha iyi olduğu belirtilmiştir.

Singh ve Kaur [6] çalışmasında tasarlanan algoritmanın sonuçlarına göre, küme içindeki noktalar ile küme merkezi arasındaki uzaklıklar toplamının küçüldüğü, buna bağlı olarak kümelerin toplam karesel hatasının daha küçük olduğu belirtilmiş, önerilen algoritmanın standart K-ortalama algoritmasından daha verimli olduğu vurgulanmıştır.

Agha ve Ashour [7], çalışmasında geliştirilen ElAgha initialization isimli algoritmada başlangıç merkezlerinin seçiminde güdümlü rasgele teknik (guided random technique) yöntemi kullanılmıştır. Önerilen algoritmanın, özellikle çok boyutlu ve karmaşık veri setlerinde kümeleme kalitesinin rasgele seçim yöntemine göre daha başarılı olduğu belirtilmiştir.
Bhardwaj ve Verma [8] çalışmasında; K-ortalama algoritmasının başlangıç parametreleri Geri Yayılım (Backpropagation) algoritması ile belirlendikten sonra standart K-ortalama uygulaması ile oluşturulan hibrit model küme sonuçlarının daha iyi olduğu vurgulanmıştır.
Oyana [9] çalışmasında k-d ağaç (k-d-tree) ve K-ortalama algoritmasını içeren FES-k-means isimli algoritma önerilmiştir. Önerilen ve standart K-ortalama algoritmaların küme sonuçlarının benzerlik gösterdiği, ancak hibrit modelin kümeleme işlem süresini azaltması önerilen algoritmanın avantajı öne çıkarılmıştır.

3. K-ortalama Kümeleme Algoritması

K-ortalama algoritması, Mac Queen tarafından 1967 yılında geliştirilen, yaygın kullanılan gözetimsiz öğrenme yöntemlerinden biridir. Yöntem en uygun küme sonucuna ulaşıncaya kadar tekrarlanan ve sürekli olarak kümelerin yenilendiği döngüsel bir algoritma olup benzer kapsamdaki algoritmaların temelini oluşturur [11,12]. Algoritma aynı zamanda diğer algoritmalarla beraber hibrit metot olarak farklı disiplinlerdeki alanlarda kullanılmaktadır [13].

K-ortalama algoritmasının adımları [14];

1. Veri seti X = (x_1, x_2, …,x_N), N elemanlı, küme sayısı k olmak üzere; k tane küme merkezi veri setinden (c_1, c_2, …, c_k) olarak rastgele seçilir.
2. Her noktanın seçilen merkez noktalara olan uzaklığı/benzerliği hesaplanır, veri noktaları hesaplanan uzaklıklara göre seçilen k tane merkezden kendine en yakın olan kümeye atanır.
3. Oluşan kümelerdeki noktaların ortalama değeri yeni küme merkezi olarak belirlenir.
4. Küme merkez noktaları değişmeyinceye kadar 2. ve 3. adımlar tekrarlanır.

K-ortalama kümeleme sonuçlarının değerlendirilmesinde genel olarak Toplam Karesel Hata (Summed Squared Error) (SSE) değeri kullanılır. Başarılı kümeleme sonuçlarının ölçütü, küçük SSE'leri en az küme sayısı ile elde etmektir, SSE Eşitlik 1'e göre bulunur [15].

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2 \ (m_i, x) \qquad (1)$$

x: *C_i kümesinde bulunan bir nesne,* ***m_i :*** *C_i kümesinin merkez noktası,* ***$dist^2(m_i, x)$:*** *C_i kümesinin m_i ile her bir x nesnesi arasındaki uzaklık ,* ***k****:küme sayısı.*

K-ortalama algoritmasının dezavantajları:

K-ortalama algoritmasında küme sayısı ve başlangıç merkezleri algoritmanın parametreleridir. Farklı parametre değerleriyle oluşturulan sonuçların değişkenlik göstermesi kümeleme sonuçlarının kalitesi açısından dezavantajdır [16]. Optimal kümelerin oluşturulması ancak, rasgele seçilen farklı başlangıç merkezleri ile elde edilen sonuçların alanında deneyimli uzmanlar tarafından gözlemsel değerlendirilmesiyle mümkündür. Bu durum özellikle büyük veri setinde analiz sürecini önemli derecede zorlaştırmakta ve küme sonuçlarının tutarsız olma ihtimalini artırmaktadır [17]. Algoritmanın başlangıç merkezlerinin rasgele belirlenmesi makul bir seçim değildir [18].

Algoritmanın dezavantajına çözüm olabilecek K-ortalama++ ve Canopy algoritmalar geliştirilmiştir.

K-ortalama++ Kümeleme Algoritması: K-ortalama++, 2007 yılında David Arthur ve Sergei Vassilvitskii tarafından geliştirilmiş K-ortalama tabanlı algoritmadır. Algoritmanın ilk aşamasında merkezler belirlenir, ikinci aşamada standart K-ortalama algoritması uygulanır. Algoritma ilk küme merkezini rasgele seçer ve diğer küme merkezler, birinci küme merkezini referans alarak ve olasılık dağılımı ile belirlenir [14].

ISSN 2422-8702 (Online), **DOI:** 10.7176/JSTR/7-01-05
Vol.7, No.1, 2021

Canopy Kümeleme Algoritması: Canopy kümeleme algoritması başlangıç merkezlerinin seçim problemine çözüm olarak kullanılan etkin algoritmalardan biridir. Algoritma 2000 yılı ACM SIGKDD konferansında McCallum, Nigam ve Ungar tarafından tanıtılmıştır. Algoritma iki aşamalı olarak tasarlanmıştır. Algoritmanın parametresi olan iki uzaklık eşik değerleri T_1 ve T_2 (T_1>T_2) belirlenir. Bu eşik değerlerine göre veri dizileri arasındaki basit uzaklık metrikleri kullanarak taslak kümeler oluşturulur. Daha sonra K-ortalama algoritması kullanarak oluşturulan taslak kümelerin iyileştirilmesi sağlanır [19].

4. Küme Geçerlilik İndeksi

Kümeleme analizinde farklı algoritmalar veya başlangıç parametreleri, değişken küme sonuçlarını ortaya çıkarabilir. Küme geçerlilik indeksi, kümeleme analizi sonucu elde edilen yapının yeterliliğini nesnel olarak ölçer. Çalışmada, dışsal (external) indekslerden Düzeltilmiş Rand indeks ölçütü kullanılmıştır.

Düzeltilmiş Rand Indeks (DRI): Özellikle nesnelerin dengesiz dağıldığı gruplarda, küme sayıları farklı olan veri setlerinde yaygın olarak tercih edilen küme geçerliliği değerlendirme indeksidir. Grup içindeki her olası nesne çifti için aynı kümede olup olmadıklarını değerlendirerek sınıfların doğru ayrılabildiği hakkında bilgi veren ölçüttür [20].
S, n tane veri nesnesi içeren veri seti olsun. S veri setinden alınmış, kümeleme algoritmasından bağımsız olarak önceden sınıflandırılmış veri seti V, U veri seti ise kümeleme algoritmasından elde edilen kümelenmiş bir bölünme. U ile V grupları karşılaştırılarak Düzeltilmiş Rand Indeks kriteri ile değerlendirildiğinde;

S veri setine ait xi, xj çiftleri V ve U bölünmelerindeki atamalarına bağlı olarak dört ayrı sonuç elde edilebilir.

a: x_i, x_j U bölünmesinde aynı kümede ve V'de aynı kategoride bulunması
b: x_i, x_j U bölünmesinde aynı kümede fakat V'de ise farklı kategoride bulunması
c: x_i, x_j U bölünmesinde farklı kümede fakat V'de ise aynı kategoride bulunması
d: x_i, x_j U bölünmesinde farklı kümede ve V'de farklı kategoride bulunması

olmak üzere; DRI [-1, 1] aralığında değer alabilir, ve Eşitlik 2'deki gibi hesaplanır.

$$\mathbf{DRI} = \frac{\binom{n}{2}(\mathrm{a}+\mathrm{d}) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]} \quad (2)$$

5. Önerilen NAMGY (Noktalar Arası Mesafe ve Gözlemlerin Yoğunluğu) Algoritması

NAMGY; K-ortalama başlangıç parametrelerini veri setindeki nesnelerin özelliklerine göre objektif yöntemlerle belirleyen K-ortalama tabanlı bir algoritmadır. Bu makalede NAMGY algoritmasının içerdiği iki parametre için geliştirilen metotlardan yalnızca başlangıç merkezlerini belirleyen yöntem açıklanarak uygulaması yapılacaktır.
NAMGY algoritmasının içerdiği metotla küme sayısı kadar uygun başlangıç merkezleri belirlenir, seçilen başlangıç merkezleri ile standart K-ortalama algoritması uygulanarak kümeler oluşturulur.

5.1. Başlangıç Merkez Seçim Metodundaki Yaklaşım

Yeterli bir kümelenme yapısı, oluşturulan sonuçların verideki doğru bilgilerle örtüşmesini veya verideki gizlenmiş örüntülerin ortaya çıkarma başarısını gösterir. Kümeleme analizinin teorisi uzaklık veya benzerlik ölçüleri kullanılarak nesnelerin benzerliklerini veya farklılıklarını ortaya koymaktır. Bu teoriye göre kümeleme analizinde grup içindeki noktalar arasındaki uzaklıkların küçük (küme içi homojenlik), gruplar arası noktalar arasındaki uzaklıkların büyük olması (kümeler arası heterojenlik) nitelikli kümeleme sonuçları için bir ölçüttür. Farklı yapılardaki veri setlerinde nesneler arasındaki uzaklıkların ve yoğunlukların değişkenlik göstermesi başlangıç merkez seçimini etkileyen faktördür. NAMGY algoritması başlangıç merkez seçiminde nesnelerin nicelik değerlerini kullanan yöntemi içeren algoritmadır. Bu özellik; başlangıç merkezlerinin uygun ve objektif seçilmesine katkı sağlayarak standart K-ortalama algoritmasının dezavantajları için bir çözüm olmaktadır.

ISSN 2422-8702 (Online), **DOI:** 10.7176/JSTR/7-01-05
Vol.7, No.1, 2021

Veri setindeki yapısal çeşitliliklerden dolayı nesneler arasındaki uzaklıkların hesaplanmasında farklı uzaklık ölçütleri kullanılabilir. Çalışmada kullanılan Öklid uzaklık ölçütü Eşitlik 3'teki gibi tanımlanır [21].

$$\boldsymbol{d(X_m, X_j)} = \sqrt{\sum_{i=1}^{n} (\boldsymbol{X_{mi}} - \boldsymbol{X_{ji}})^2} \quad (3)$$

d (X$_m$, X$_j$) *: X_m noktası ile X_j noktası arasındaki öklit uzaklığı,* ***n:*** *Veri setindeki boyut sayısı.*

Yöntem için kullanılan terimlerin açıklamaları ve eşitlikleri aşağıda verilmiştir.

Genel Ortalama Uzaklık (GOU): Veri setindeki noktalar arasındaki uzaklıkların ortalaması Eşitlik 4'e göre bulunur.

$$\mathbf{GOU} = [\sum_{i=1}^{N}\sum_{j=1}^{N} \boldsymbol{d(X_i, X_j)}\ ]/((\mathbf{N} * \mathbf{N}) - \mathbf{N}), \boldsymbol{i \neq j} \quad (4)$$

GOU: *Genel Ortalama Uzaklık,* ***d(X$_i$,X$_j$):*** *X_i noktası ile X_j noktası arasındaki uzaklık,* ***N:*** *Toplam kayıt sayısı*

Komşu Sayısı (KS): Her noktanın kendisine GOU'dan küçük eşit olan noktaların sayısı Eşitlik 5'e göre bulunur.

$$\mathbf{d(X_i, X_j) \leq GOU}\ \textit{ise}\ X_i\ \textit{ile}\ X_j\ \textit{noktaları komşu} \quad \mathbf{(5)}$$

Komşu Nokta Uzaklık Ortalaması (KUO): Komşu noktalar arasındaki uzaklıkların ortalaması Eşitlik 6'ya göre bulunur.

$$\boldsymbol{X_i}\mathbf{KUO} = [\sum_{j=1}^{KS_i} \boldsymbol{d(X_i, X_j)}] \ / \ \boldsymbol{KS_i}, \boldsymbol{i} = \mathbf{1, 2, \ldots N} \quad (6)$$

X$_i$KUO : *X_i' nin komşu noktalarına uzaklıklar ortalaması,* ***KS$_i$:*** *X_i' nin toplam komşu nokta sayısı*
d(X$_i$,X$_j$) : *X_i noktası ile kendisine komşu olan X_j noktası arasındaki uzaklık,* ***N :*** *Toplam kayıt sayısı*

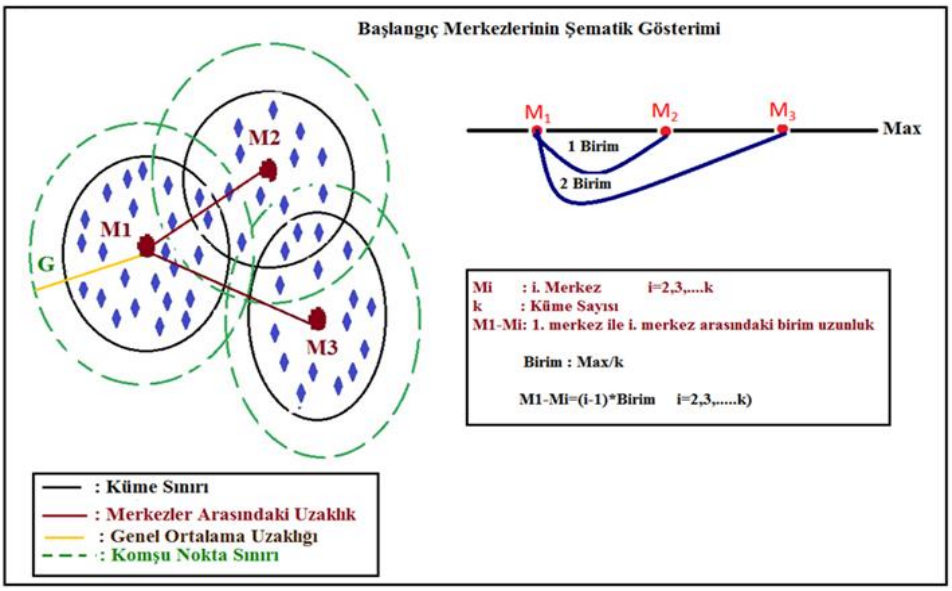
Birinci Merkez (M$_1$): Bu merkezin seçiminde; komşu nokta sayısı (KS_i), komşu noktalarına uzaklıklar ortalaması (X_iKUO) ise ***KS$_i$ / X$_i$KUO*** oranı ve ***(KS$_i$)*** en büyük değerler kullanılır.

Maksimum Uzaklık (Max): Birinci merkez ile kendisine en uzak nokta arasındaki uzaklık.

Birim Uzaklık (Unit): Birinci merkeze en uzak noktanın uzaklığı (maxuz) küme sayısına bölünerek bulunur. ***(Unit=maxuz / k,*** ***k:*** *küme sayısı).*

Bu tanımlara göre; küme sayısı (k) 3 olarak alındığında başlangıç merkezlerini belirleme yönteminin şematik gösterimi Şekil 1'deki gibidir.

ISSN 2422-8702 (Online), **DOI:** 10.7176/JSTR/7-01-05
Vol.7, No.1, 2021



Şekil 1: Başlangıç Merkez Seçim Metodunun Şematik Gösterimi

5.2. NAMGY Algoritması (Başlangıç Merkez Seçimi)

Önceden belirlenmiş küme sayısı (k) kadar başlangıç merkezi bulunur, K-ortalama algoritması seçilen merkezler ve küme sayısı parametre değerleriyle uygulanır. Algoritmanın aşamaları aşağıdaki gibidir.

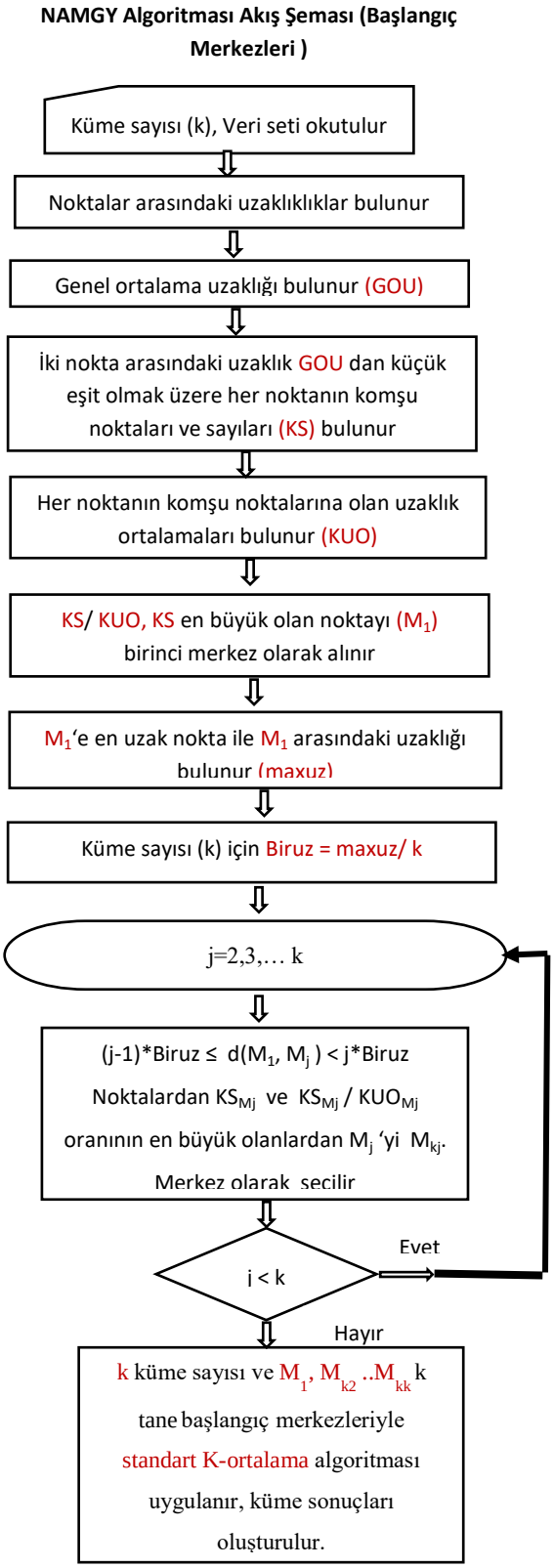
1. Küme sayısı ve veri seti okutulur.
2. Veri kümesindeki her noktanın diğer noktalara olan uzaklıkları bulunur.
3. Veri kümesindeki noktalar arasındaki Genel Ortalama Uzaklık (GOU) Eşitlik 4'e göre hesaplanır.
4. Her noktanın, GOU'dan küçük eşit uzaklıktaki noktaların sayıları komşu nokta (KS) Eşitlik 5'e göre bulunur.
5. Her noktanın kendi komşu noktalarına olan uzaklıkları veri yapısına uygun uzaklık formüllerine göre bulunur.
6. Her noktanın kendi komşu noktalara olan uzaklıklar ortalaması KUO Eşitlik 6'ya göre bulunur.
7. Birinci merkez (M_1) noktası, komşu nokta sayısı (*KS*), komşu noktalarına uzaklıklar ortalaması (KUO) olmak üzere, **(*KS* / KUO)** ve **(*KS*)**'nin en büyük olma koşulunu sağlayan noktalardan seçilir.
8. Birinci merkez (M_1) ile kendine en uzak nokta arasındaki uzaklık (**maxuz)** bulunur.
9. Küme sayısı (k), j=2, $d(M_1,M_{kj})$: M_1 ile M_{kj} noktalar arasındaki uzaklık hesaplanır. (k:küme sayısı, j:merkez sırası, M_{kj}: k küme sayılı j.merkez)
10. Küme sayısı (k) için ***Biruz = maxuz / k***
11. Diğer merkezlerin seçimi için;

$$(j-1)*Biruz \leq d(M_1, M_j) < j*Biruz \qquad (7)$$

Eşitlik 7'deki koşulu sağlayan noktalardan; komşu nokta sayısı (KS_i), komşu noktalarına uzaklıklar ortalaması (X_iKUO) ise **(KS_i / X_iKUO)** ve **(KS_i)** değerlerinin en büyük olduğu nokta M_j'yi k sayılı kümenin j. merkez M_{kj} olarak seçilir.

12. Merkez seçimi j < k ise j=j+1, **11. Adıma** gidilir.
13. k küme sayısı ve M_1, ..M_{kk} başlangıç merkezleriyle standart K-ortalama algoritması uygulanır.

NAMGY algoritması başlangıç merkezleri ile ilgili yöntemin akış şeması Şekil 2'de verilmiştir.



Şekil 1. NAMGY Algoritmasının Akış Şeması (Başlangıç Merkezi)

6. Yöntem ve Materyal

Bilimsel çalışmalarda yaygın olarak kullanılan WEKA, çalışmamızda referans program olarak kullanılmıştır. WEKA, Yeni Zelanda'nın Waikato Üniversitesi tarafınca geliştirilen, açık kaynak kodlu modüler bir veri madenciliği programıdır [22]. WEKA programında K-ortalama algoritmasının başlangıç merkezlerinin seçimi için Random, K-means++, Canopy olmak üzere üç farklı seçim yöntem sunulmuştur. Aynı veri setleri üzerinde NAMGY(K-ortalama) ve WEKA(K-ortalama) algoritmalarının uygulaması yapılmıştır.

6.1. Veri Setleri

NAMGY ve WEKA programlarının içerdiği K-ortalama algoritması, bilimsel çalışmalarda sıkça kullanılan İris, Yeast ve Segment-Challenge veri setleri üzerinde uygulanmıştır. Ayrıca iki algoritmanın çalışma prensibi açısından karşılaştırılmasında Çukurova Üniversitesi Balcalı Hastanesi Hastane Bilgi Yönetim Sisteminden alınan VitaminB12 veri seti üzerinde algoritmalar uygulanarak sonuçlar oluşturulmuştur. Veri setleri üzerinde normalizasyon işlemi yapılarak veri setlerindeki nesneler [0,1] aralığında değerlere dönüştürülmüştür. Nesneler arasındaki uzaklık ölçümlerinde Öklid uzaklık ölçütü kullanılmıştır. Veri setlerinin genel özellikleri Tablo.1'de verilmiştir.

Tablo 1. Veri Seti Özellikleri

Veri Seti	Kayıt Sayısı (N)	Küme Sayısı	Öznitelik Sayısı
Iris	150	3	4
Yeast	1484	10	8
Segment- Challenge	1500	7	19
VitaminB12	10685	5	4

7. Bulgular

Algoritmaların veri setleri üzerine uygulama sonuçları üç başlık altında sunulmuştur.

Algoritmanın çalışma prensibi; NAMGY algoritmasının analiz sürecindeki işlemler açısından standart K-ortalama algoritmasıyla karşılaştırılması yapılmıştır.

Başlangıç merkezlerinin küme sonuçlarına Etkisi; Farklı başlangıç merkezleri ile K-ortalama algoritmasının oluşturduğu değişken küme sonuçları gözlemsel olarak değerlendirilerek optimal kümeleri belirlenir. Farklı SSE değerlerine göre oluşturulan kümelerin yorumlanması karşılaştırılmıştır.

Kümeleme performanslarının değerlendirilmesi; NAMGY algoritması ve standart K-ortalama algoritmasının küme sonuçları küme geçerlilik indekslerine göre değerlendirilmiştir.

7.1 Algoritmaların Çalışma Prensipleri

NAMGY(K-ortalama) ve referans olarak kullanılan WEKA(K-ortalama) algoritması (başlangıç merkez seçimi Random seçeneği) VitaminB12 veri seti üzerinde uygulanarak işlem süreçleri çerçevesinde karşılaştırılmıştır.

Çukurova Üniversitesi Balcalı Hastanesi 2016 yılında hastaneye gelen 10685 hastaya ait; anabilim dalı, cinsiyet, yaş, vitaminB12 verilerden oluşan VitaminB12 veri seti üzerinde WEKA (K-ortalama) ve NAMGY (K-ortalama) algoritmaları uygulanmıştır. Ön çalışmalar yapılarak VitaminB12 veri seti için uygun küme sayısı 5 olarak belirlenmiştir.

WEKA (K-ortalama) Programının Uygulanması; Rasgele seçilmiş 18 farklı seed değeri ile oluşturulan küme sonuçları irdelendiğinde; 18 farklı seed ve iterasyon sayısına karşın 6 farklı SSE değeri bulunmuştur. En küçük SSE değeri ve en küçük iterasyon sayısını veren seed değeri gözlemsel olarak uygun başlangıç kabul edilip optimal kümeler oluşturulmuştur. (SSE: 489.67, iterasyon sayısı:14, seed:10).

NAMGY (K-ortalama) Programın Uygulanması; Öncelikle, başlangıç merkezleri NAMGY algoritmasını içeren arayüz programı ile uygun başlangıç merkezler belirlenmiştir. Algoritmik parametre değerlerinden k=5 (küme sayısı) ve seçilen başlangıç merkezleriyle NAMGY(K-ortalama) algoritması VitaminB12 veri seti üzerinde uygulanarak kümeleme işlemi tamamlanmıştır.

İki algoritmanın uygulama sonuçlarına göre çalışma prensibi açısından farlılıklar gözlenmiştir. NAMGY (K-ortalama) algoritmasında başlangıç merkezleri nesnelerin özellikleri dikkatte alınarak algoritmanın içerdiği yöntemle ve alan deneyimi gerektirmeden programın bir kez çalıştırılarak optimal

küme sonuçları oluşturulmuştur. Oysa WEKA (K-ortalama) uygulanmasında uygun başlangıç merkezlerinin belirlenmesi, farklı seed değerleriyle 18 kez çalıştırılarak gözlemsel olarak optimal küme sonuçları oluşturulmuştur. Standart K-ortalama algoritmada geçerli küme sonuçları oluşturmak için tekrarlı denemelerin yapılmasının gereksinim olduğu görülmüştür. İki algoritmanın karşılaştırmalı sonuçları Tablo 2'de verilmiştir.

Table 2. İki Algoritmanın İşlem Sürecindeki Değişken Değerleri (VitaminB12 veri seti)

Algoritma	Küme Sayısı (k)	Algoritmanın Çalıştırma Sayısı	Seed Değeri	İterasyon Sayısı	SSE
WEKA (K-ortalama)	5	18	10	14	489.67
			7654	17	
			2222	23	
			980	26	
			1453	30	
			7690	32	
			5871	13	522.64
			9870	16	
			8710	23	
			356	16	558.26
			5000	20	
			3335	12	721.49
			4908	28	
			54	30	
			1870	44	746.19
			7356	14	
			367	27	755.78
			2	16	
NAMGY (K-ortalama)	5	1	__	12	489.66

WEKA(K-ortalama) ve NAMGY(K-ortalama) algoritmalarının optimal kümeleme sonuçlarının örtüştüğü görülmüştür. Ancak kümeleme sürecine etki eden başlangıç merkezlerinin seçim yöntemine bağlı olarak değişen tekrar (iterasyon) ve seed sayıları açısından farklılık göstermiştir. Algoritmaların optimal küme sonuçlarına göre parametre değerleri Tablo.3'te, çalışma prensipleri açısından farklılıklar Tablo.4'te verilmiştir.

Table 3. Algoritmaların Optimal Küme Sonuçlarına Göre Parametre Değerleri (VitaminB12 veri seti)

K-ortalama	Küme Sayısı	İterasyon Sayısı	SSE	Rasgele Seçilen Seed Sayısı
WEKA	5	14 (En küçük)	489.67	18
NAMGY	5	12	489.66	0

ISSN 2422-8702 (Online), **DOI:** 10.7176/JSTR/7-01-05
Vol.7, No.1, 2021

Table 4. Algoritmaların Çalışma Prensiplerindeki Farklılıklar

İşlevsellik	**NAMGY (K-ortalama)**	**WEKA (K-ortalama)**
Başlangıç küme merkezlerin seçimi	Algoritma	Rasgele
Optimal kümeleme İçin uygulamayı çalıştırma sayısı	1	Farklı başlangıç değerleri çalıştırılarak gözlemsel belirlenir.
Kullanıcı Dostu	Evet	Hayır

WEKA(K-ortalama) programında seçenek olarak sunulan K-ortalama++, Canopy yöntemleri ve NAMGY standart K-ortalama algoritmasının dezavantajı olan başlangıç merkezlerini belirleme problemine çözüm olabilecek algoritmalardır. Küme merkezlerinin seçiminde benzer yaklaşımlar kullanılsa da algoritmaların yöntemleri arasında farklılıklar gözlenmektedir. K-ortalama++, Canopy algoritmalarında merkezler bir yöntemle belirlenmesine rağmen rasgele seçilen parametre değerlerine ihtiyaç duyulmaktadır, oysa NAMGY analiz sürecinde parametre gereksinimi yoktur. Yöntemlerin karşılaştırma sonuçları Tablo 5'te gösterilmiştir.

Table 5. K-ortalama++, Canopy ve NAMGY Algoritmaların Merkez Seçimlerindeki Yaklaşımı

Algoritma	**İlk Merkez Seçimi**	**Diğer Merkezlerin Seçimi**	**Algoritmanın Parametresi**
K-Ortalama++	Seed Değeri	Algoritma Rasgele	SEED
Canopy	Seed Değeri	Algoritma Rasgele	SEED (T_1,T_2)
NAMGY	Algoritma	Algoritma	YOK

7.2 Başlangıç Merkezlerinin Küme Sonuçlarına Etkisi

K-ortalama yönteminde küme sonuçlarının parametre değerlerine göre değişkenlik göstermesi algoritmanın zayıf noktasıdır. Geçerli küme sonuçları için başlangıç merkezlerinin dikkatli seçilmesi kaçınılmazdır. Bu uygulamada, aynı veri seti üzerinde algoritmaların oluşturduğu küme sonuçlarının tutarlılığını gösteren çıkarımlar yapılmıştır.

Her iki algoritmanın küme sonuçları niteliksel olarak karşılaştırılarak, başlangıç merkez seçiminin küme sonuçlarına etkisi araştırılmıştır. Weka (K-ortalama) algoritmasının VitaminB12 veri seti uygulamasında optimal küme sonucunu belirleme sürecinde rasgele seçilen başlangıç değerleriyle hesaplanan SSE değerlerinin (SSE=755.78, SSE=742.99, SSE= 489.667) oluşturduğu küme sonuçları ve NAMGY(K-ortalama) algoritmasının hesapladığı 489.666 SSE değerinin küme sonucu örnek alınarak, VitaminB12 değişkeninin cinsiyet ve yaş üzerine etkisini gösteren çıkarımlar yapılmıştır. WEKA(K-ortalama) algoritmasının rasgele seçilmiş 18 seed değerleri ile hesaplanan 6 farklı SSE'den büyük olan iki SSE'nin küme sonuçlarının değerlendirilmesi Tablo 6'da verilmiştir. Her iki algoritma için küçük SSE koşulunu sağlayan optimal küme sonuçlarının değerlendirilmesi Tablo 7'de sunulmuştur.

(Çalışmada vitaminB12 değerleri hastane bilgi yönetim sistemindeki sınır değerlere göre Alt (< 127), Normal (≤ 127 ve < 555) , Üst (≤ 555) olmak üzere üç düzeyde sınıflandırılmıştır).

ISSN 2422-8702 (Online), **DOI:** 10.7176/JSTR/7-01-05
Vol.7, No.1, 2021

Tablo 6. K-ortalama Algoritmasında Farklı Başlangıç Merkezleriyle Oluşturulan Küme Sonuçları

Küme Sonuçların Yorumu	**WEKA(K-ortalama)**	
	SSE=755.78	***SSE=742.99***
Çıkarım-1	VitaminB12 düzey oranları her yaş grubundaki kadınlarda benzerdir	VitaminB12 düzey oranları her yaş grubundaki erkeklerde benzerdir
Çıkarım-2	VitaminB12 eksikliği oranı erkek çocuklarında en düşüktür	VitaminB12 eksikliği oranı kız çocuklarında en düşüktür
Çıkarım-3	VitaminB12 normal düzey oranı genç yaş grubundaki erkeklerde en yüksektir	VitaminB12 üst düzey oranı orta yaş grubu kadınlarda en yüksektir

Tablo 7. WEKA, NAMGY K-ortalama Algoritmasının Optimal Küme Sonuçlarına Göre Değerlendirilmesi

Küme Sonuçları Yorumu	**K-Ortalama**	
	WEKA	***NAMGY***
	SSE=489.667	***SSE=489.666***
Çıkarım-1	VitaminB12 normal düzey oranlarında orta yaş grubu hastalarda cinsiyetlere göre farklılık yoktur.	VitaminB12 normal düzey oranlarında orta yaş grubu hastalarda cinsiyetlere göre farklılık yoktur.
Çıkarım-2	VitaminB12 eksikliği oranı orta yaş grubu erkeklerde en yüksektir.	VitaminB12 eksikliği oranı orta yaş grubu erkeklerde en yüksektir.
Çıkarım-3	VitaminB12 üst düzey oranı orta yaş grubu kadınlarda en yüksektir.	VitaminB12 üst düzey oranı orta yaş grubu kadınlarda en yüksektir.
Çıkarım-4	VitaminB12 eksikliği kız çocuklarında en düşüktür.	VitaminB12 eksikliği kız çocuklarında en düşüktür.

WEKA(K-ortalama) algoritmasının farklı başlangıç merkezleri ile oluşturulmuş SSE'lere ait küme sonuçlarına göre yapılan çıkarımlarda farklılık gözlenmiştir. Aynı veri setinden oluşturulan küme sonuç bilgilerinin birbiriyle tutarsız olması bilimsel çalışmaların güvenirliliğini azaltmaktadır. Tablo.6 ve Tablo.7 'den görüldüğü gibi; K-ortalama algoritmasının başlangıç merkezlerinin rasgele değil, veri setindeki noktaların özelliklerini kullanan yöntemle seçilmesi küme sonuçlarını önemli derecede etkilemekte ve geçerliliği artırmaktadır. Bu durum K-ortalama algoritması için başlangıç merkezlerinin dikkatli seçilmesi ve küme sonuçlarının alanında deneyimli uzmanlar tarafından yorumlanması gerekliliğini ortaya koymaktadır.
NAMGY(K-ortalama) algoritması kümeleme işlem sürecinde gerek çalışma prensibi açısından, gerekse küme sonuçlarının tutarlılığı açısından daha verimli olma beklentisinin çalışma sonuçlarıyla desteklendiği görülmektedir.

7.2. Algoritmaların Küme Geçerliliği

NAMGY ve WEKA K-ortalama algoritmaları bilimsel çalışmalarda kabul görmüş sınıflandırılmış Iris, Yeast, Segment Challenge normalize yapılmış veri setleri üzerinde uygulanarak oluşturulan optimal kümeleme sonuçları belirlenmiştir. NAMGY algoritmasında her üç veri seti için DRI değerleri daha yüksek, aynı zamanda daha düşük iterasyon ile daha düşük SEE değerleri bulunduğu görülmüştür. Üç performans kriterine göre NAMGY algoritmasının daha avantajlı olduğu söylenebilir, sonuçlar Tablo 8'de gösterilmiştir.

Tablo 8. NAMGY ve WEKA K-ortalama Algoritmaların Küme Geçerlilik Kriterlerine Göre Sonuçları

Veri Seti	Performans Kriteri	NAMGY (K-ortalama)	WEKA (K-ortalama)
IRIS	DRI İndeks	0.701	0.690
	SSE	7.130	7.140
	İterasyon Sayısı	2	3
YEAST	DRI İndeks	0.014	0.013
	SSE	58.965	69.946
	İterasyon Sayısı	22	35
SEGMENT CHALLENGE	DRI İndeks	0.414	0.396
	SSE	299.341	327.990
	İterasyon Sayısı	7	8

8. Sonuç ve Tartışma

K-ortalama kümeleme algoritması bilimsel çalışmalarda yaygın olarak kullanılmaktadır. Kümeleme sonuçlarını önemli derecede etkileyen başlangıç parametrelerinin kullanıcı tarafından gözlemsel olarak belirlenmesi algoritmanın dezavantajıdır.Bu çalışmada, tarafımızdan geliştirilen, hem başlangıç merkezlerini seçim hem de küme sayısını tespit yöntemlerini içeren, K-ortalama tabanlı algoritma (NAMGY)'nın, merkez belirleme parçası uygulamalarla desteklenerek tanıtılmaktadır. Aynı amaçla tasarlanan algoritmalar, genel olarak, başlangıç merkezlerinin seçiminde nesneler arasındaki uzaklıkları veya nesnelerin yoğun bulunduğu bölge özelliklerini temel alan farklı yaklaşımları içerirler, oysa NAMGY algoritmasında uygun merkezler, nesnelerin her iki özelliğine göre algoritmayla belirlenmektedir. Algoritmada merkezlerin belirlendiği ilk bölümde, veri kümesindeki her noktanın diğer tüm noktalar arasındaki uzaklıkların kullanılması özellikle büyük veri setleri için algoritmanın zayıf noktası olabilir. Ancak algoritmanın arayüz programında çoklu iş (multitasking) tekniği kullanılarak problem giderilebilir.

NAMGY ile önerilen algoritmalar arasındaki başlangıç merkezlerini belirlemede önemli yöntem farklılıkları vardır.

Diğer algoritmaların kullandığı yöntemlerde, eşik değerinin en küçük ve en büyük uzaklıklar ile belirlenmesi daha küçük/büyük eşik değeri bulunmasına neden olabilir. Bu durum, özellikle algoritmanın uç noktalarının çok olduğu bir veri setinde tekrar (iterasyon) sayısını büyütebileceği gibi, küme içi homojenliğin veya kümeler arası heterojenliğin bozulma olasılığını artırabilir. NAMGY algoritmasında, eşik değerinin noktalar arasındaki uzaklıklar ortalamasına dayalı belirlenmesi daha uygun merkez seçimine katkı sağlayabilir.

Merkez seçiminde, noktalar arasındaki uzaklıklar toplamının büyüklüklerine göre sıralanıp, küme sayısı kadar ayrılan parçalardan ilk noktanın sistematik olarak merkez belirlenmesi farklı yapılardaki veri setleri için uygun olmayabilir. Özellikle heterojen bir veri setinde, seçilen noktanın aykırı nokta olması durumunda, iterasyon sayısı artabilir. NAMGY algoritmasında merkez seçiminde uzaklıklarla beraber noktaların komşu noktalarla ilgili niceliksel verilerinin değerlendirilmesi nedeni ile algoritmanın heterojen veri seti içinde geçerli bir yöntem olduğu söylenebilir.

K-ortalama algoritmasının başlangıç merkezlerinin seçimi için geliştirilen K-ortalama++, Canopy algoritmaları probleme kısmen çözüm olsa da, NAMGY ile çalışma sistemleri açısından da aşağıdaki farklılıklar gözlenmiştir.

NAMGY algoritmasında başlangıç merkez seçiminin kullanıcıdan bağımsız, nesnelerin niceliksel değerlerine göre uygun seçimlerin algoritma tarafından saptanması bir avantajdır.

NAMGY ile K-ortalama++ algoritmaları başlangıç merkezlerini belirleme yaklaşımlarında benzerlik göstermektedir. Ancak K-ortalama++ olabilecek merkezleri metotla belirlemesine rağmen kullanılacak merkezleri rasgele seçmektedir. NAMGY algoritmasında bu işlem veri setinin yapısına bağlı bir yöntemle yapılmaktadır.

NAMGY ve Canopy algoritmaları başlangıç merkezlerini belirleme yöntemlerinde eşik uzaklık değeri kullanması algoritmaların benzerliğidir. Ancak Canopy algoritmasının eşik değerleri (T1, T2) kullanıcı tarafından çok büyük veya çok küçük seçilmesi durumunda birinci aşamadaki taslak kümelerin

oluşturulması başarısız olabilir. Dolayısıyla parametre için uygun değerlerin objektif olarak belirlenmesi gerekir. NAMGY algoritmasında, eşik uzaklık değerinin verilere dayanarak belirlenmesi optimal küme sonuçlarının oluşmasına destek sağlayarak algoritmanın başarısını artırabilir.
NAMGY ve standart K-ortalama algoritmaları Iris, Yeast ve Segment-Challenge veri setleri uygulamalar farklı açılardan değerlendirilmiştir. Analiz süreçlerindeki farklılıklar ve küme geçerliliğini gösteren ölçütlere göre karşılaştırıldığında, NAMGY algoritmasının daha etkin analiz süreci ve kümeleme sonuçları oluşturduğu görülmüştür.
NAMGY algoritmasında başlangıç merkezlerinin seçiminde uygulanan yöntemin algoritmanın tekrar sayısını azaltması, büyük veri setleri için önemli bir avantajdır. Aynı zamanda başlangıç merkezlerin gözlemsel olarak değil, bir yöntemle belirlenmesi objektif sonuçların oluşturulmasına destek sağlayarak bilimsel çalışmaların tutarlılığını artırabilir.
Önerdiğimiz algoritmada parametre seçiminin profesyonel olarak yapılması algoritmanın kullanılabilirliğini ve kümeleme sonuçlarının güvenirliliğini artırmak açısından önemlidir.
Standart K-ortalama algoritmasının dezavantajı olan başlangıç merkezlerini belirleme problemine NAMGY bir çözüm seçeneğidir. NAMGY algoritmasının çeşitli kaynaklardan ve farklı formattaki çok boyutlu karmaşık nesnelerin oluşturduğu veri setleri üzerindeki uygulamaları ve performans araştırması sonraki çalışmaların konusu olabilir.

Kaynaklar

[1] Mirkin B. Clustering for Data Mining. U.S: Chapman & Hall/CRC Taylor & Francis Group. 2005.

[2] Dalhatu K, Sim A.T.H. Density base k-Means Cluster Centroid Initialization Algorithm. International Journal of Computer Applications, 137 (11):49-51, 2016.

[3] Çolak B, Durdağ Z, Erdoğmuş P. K-Means Algoritması İle Otomatik Kümeleme. El-Cezerî Fen ve Mühendislik Dergisi. 3(2):315-323, 2016.

[4] Godara A, Sharma V. Improvement Of Initial Centroids in K Means Clustering Algorithm. IJARIIE-ISSN(O), 2(2):2395-4396, 2016.

[5] Kedar B, Sawant K. B. Efficient Determination of Clusters in K-means Algorithm Using Neighborhood Distance, International Journal of Emerging Engineering Research and Technology, 2349(4409):22-27, 2015.

[6] Singh H, Kaur K. New Method for Finding Initial Cluster Centroids in K-means Algorithm, International Journal of Computer Applications, 74(6):27-30, 2013.

[7] Agha M, Ashour W. Efficient and Fast Initialization Algorithm for K-means Clustering, I.J. Intelligent Systems and Applications, 21(31), 2012.

[8] Bhardwaj S, Verma V. Improved K-means Clustering Algorithm Using Black Propagation Method, I J C T A, 9(11): 5169-5180, 2016.

[9] Oyana T. J. A New-Fangled FES-k -Means Clustering Algorithm for Disease Discovery and Visual Analytics, EURASIP Journal on Bioinformatics and Systems Biology, doi:10.1155/2010/746021,14 pages, 2010 .

[10] Ceylan Z, Gürsev S, Bulkan S. İki Aşamalı Kümeleme Analizi ile Bireysel Emeklilik Sektöründe Müşteri Profilinin Değerlendirilmesi, Bilişim Teknolojiler Dergisi, 10(4): 475-485, 2017.

[11] Silahtaroğlu G. Kavram ve Algoritmalarıyla Temel Veri Madenciliği, Papatya Yayıncılık, İstanbul, 2008.

[12] King R.S. Cluster Analysis And Data Mining, Mercury Learning and Information LLC, New Delhi, 2015.

[13] Arı E. S, Özköse H, Doğan A, Calp M. H. İstanbul Borsası'nda İşlem Gören Firmaların Finansal Performanslarının Kümeleme Analizi ile Değerlendirilmesi, Bilişim Teknolojiler Dergisi, 9(1): 33-39, 2016.

[14] Arthur D. Vassilvitskii S., k-means++: The Advantages of Careful Seeding, Proceeding SODA '07 Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithmS, 1027-1035, 2007.

[15] Han J. M. Kamber, J. Pei, Data Mining Concepts and Techniques, 3nd Ed, Morgan Kaufmann Publishers is an imprint of Elsevier, USA, 2012.

[16] Wierzchoń S.T, Kłopotek M.A. Modern Algorithms of Cluster Analysis, Springer International Publishing AG, Switzerland, 2018.

[17] Göral Yıldızlı M, Alparslan Z.N. K-Ortalama Kümeleme Yönteminde Başlangıç Merkezlerinin Kümeleme Sonuçlarına Etkisi, XIX. Ulusal ve II. Uluslararası Biyoistatistik Kongresi, 25-28 Ekim 2017 Antalya.

[18] B. Mirkin, Mathematical Classification And Clustering, Dordrecht Kluwer Academic Publishers, Boston, 1996.

[19] Kumar A, Ingle I.Y.S, Pande A, Dhule P. Canopy Clustering: A Review on Pre-Clustering Approach to K-Means Clustering, International Journal of Innovations & Advancement in Computer Science, 3/5: 24-28, 2014;

[20] Santos J.M, Embrechts M. On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification, In Proceedings of the 19th International Conference on Artificial Neural Networks (ICANN), Part II Lecture Notes in Computer Science, vol5769: pp 175-184. Springer, Berlin (2009).

[21] Akpınar H. Data Veri Madenciliği Veri Analizi, 1. Basım, Papatya Yayıncılık, İstanbul, 2014.

[22] The University of Waikato, WEKA V, https://www.cs.waikato.ac.nz/ml/weka/ (İnternet), 2014.