**KATEGORİK VERİLER İÇİN KÜMELEME ANALİZİ**

**Barış ERGÜL1, Arzu ALTIN YAVUZ2**

**1** Eskişehir Osmangazi Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü, Eskişehir/ TÜRKİYE,

Orcid No: 0000-0002-1811-5143

**2** Eskişehir Osmangazi Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü, Eskişehir/ TÜRKİYE

Orcid No: 0000-0002-3277-740X

**Özet**

***Amaç:*** *Kümeleme Analizi, veri matrisinde yer alan ve doğal gruplamaları kesin olarak bilinmeyen birimleri, birbiri ile benzer olan alt kümelere ayırmaya yardımcı olan yöntemler topluluğudur. Kümeleme analizi; birimleri veya değişkenler arası benzerlik ya da farklılıklara dayalı olarak hesaplanan bazı ölçülerden yararlanarak homojen gruplara bölmek amacıyla kullanılır. Kümeleme Analizi; ekonomiden psikolojiye, tıptan ziraat bilimine ve biyolojiye kadar birçok bilim dalında sınıflandırma yapmak amacıyla kullanılmaktadır. Bu çalışmanın amacı, hayvan türlerini birbirlerine olan benzerliklerine ve farklılıklarına göre, kümeleme analizi kullanılarak sınıflandırmak ve sınıflandırma performanslarını karşılaştırmaktır.*

***Yöntem:*** *Kümeleme Analizinde, veri setinde yer alan değişkenlerin ölçme düzeyleri, sıralayıcı ve sınıflayıcı olduğunda, klasik kümeleme teknikleri işlemez duruma gelir. Bu durumda, farklı benzerlik ölçütleri kullanılarak analizlere devam edilmesi gereği ortaya çıkmaktadır. Kategorik verilerde kullanılan benzerlik ölçütlerinden birisi de Jaccard benzerlik ölçüsüdür. Bu benzerlik ölçütü kullanılarak, kümeleme analizi yapılmaktadır. Diğer bir kümeleme tekniği ise, K-modes tekniğidir. Bu teknik, K-Ortalamalar tekniğinin bir uzantısı olarak karşımıza çıkmaktadır. Bu teknikte, ortalamalar yerine mod kullanılmaktadır. Kategorik verilerin kümeleme teknikleri içinde sağlam teknikler de vardır. Bunlardan en sık kullanılanı ROCK (RObust Clustering using linKs) tekniğidir. ROCK tekniği, aynı kümeden gözlemler için bağlantıların toplamı maksimize edilmesini ve farklı kümelerdeki gözlemler için bağlantıların toplamlarının en aza indirgenmesini dikkate alan fonksiyonun maksimize edilmesini sağlar.*

***Bulgular:*** *İlk aşamada, dünya üzerinde yaşayan ve türleri bilinen 20 adet hayvan 6 değişken bakımından (sıcak kanlı olma durumu, uçabilme durumları, omurgalı olma durumları, neslinin tükenme durumu, gruplar halinde yaşama durumları, saç durumu) oluşturulan kategorik veri setinden yola çıkılmıştır. Veri setinde ilgili özelliğe sahip olmayan türe 1, sahip olan türe 2 kodu verilmiştir. Sonraki aşamada, uygun küme sayısına çeşitli ölçütler göz önüne alınarak karar verilmiştir. Sonraki aşamalarda ilgili kategorik veri setinden oluşan hayvan veri seti için Jaccard, K-modes ve ROCK kümeleme teknikleri uygulanmıştır. Her bir hayvanın temsil ettiği kümelere ait sınıflar belirlenmeye çalışılmıştır.*

***Sonuç:*** *Hayvan türlerine ait sınıflandırma performansları çeşitli ölçütler bakımından karşılaştırılmıştır.*

***Anahtar Kelimeler:*** *Hayvan, K-modes, ROCK, Jaccard, Kümeleme*

**Cluster Analysis for Categorical Data**

**Abstract**

***Purpose:*** *The clustering analysis is that the natural groupings in the data matrix are a collection of methods that help separate unknown units. It is a clustering analysis that is used to divide some dimensions based on units or variables based on similarities or differences based on homogeneous groups. The clustering analysis is used in psychology from the economy to agricultural science and biology of medicine to graduate in many sciences. This study aims to classify and compare classification performance using clustering analysis according to their similarities and differences of animal species.*

***Methods:*** *However, when the variables in the data set are measurement levels, ordinal and nominal, classical clustering techniques are not operating. In this case, it is necessary to continue analyzes using different similarity criteria. One of the similarity criteria used in categorical data is the measure of Jaccard similarity. By using this similarity measure, clustering analysis is performed. Another clustering technique is the K-Modes technique. This technique is an extension of the K-means technique as an extension. This technique uses modes instead of means. There are also robust techniques in clustering techniques of categorical data. The most commonly used is the Rock (Robust Clustering Using Links) technique. The Rock technique allows the total function to maximize the sum of the connections for observations and to maximize the functions that are considered to be minimized for the observations in different sets.*

***Findings:*** *In the first stage that live on the world and the types of known 20 animals and, 6 variable (warm-blooded, can fly, vertebrate, endangered, conditions of living in groups, have hair) formed from the categorical data set. In the data set; Type 1 that do not have the relevant feature and Type 2 that the relevant feature. At the next stage, the appropriate cluster number is decided to consider a variety of criteria. In the next stages, Jaccard, K-Modes and Rock clustering techniques were applied to the animal data set consisting of the categorical data set. Classification belonging to each animal represented are tried to be determined.*

***Results:*** *The classification performances of animal species were compared in terms of various criteria.*

***Keywords:*** *Animal, K-modes, ROCK, Jaccard, Cluster*