

Kümeleme Tekniklerinin Temel Bilimlerde Kullanımı

Oğuz Akpolat^{1*}, Sinem Çağlar Odabaş², Gülçin Özevci³, Nezahat İpdeş⁴

¹Muğla S.K.Ü, Fen Fakültesi, Kimya Bölümü, Muğla, Türkiye

²Muğla S.K.Ü Fen Bilimleri Enstitüsü, Kimya AnaBilim Dalı

³Ege Üniversitesi, Nükleer Bilimler Enstitüsü, İzmir, Türkiye

⁴Muğla S.K.Ü Fen Bilimleri Enstitüsü, Çevre Bilimleri AnaBilim Dalı

*oakpolat@mu.edu.tr

Özet: Verilerin kümelenmesinin amacı heterojen olan ana kütleli homojen gruplara ayırmaktır. Kimyasal analizlerde de ham veri büyük öneme sahiptir bu değişkenler için kümeleme işleminin gerçekleştirilebilmesi değişkenler arası benzerlik ya da farklılıkların bulunmasına dayanır. Kümeleme yöntemleri, birim ya da değişkenleri uygun gruplara ayırırken grupları belirlemede izledikleri yaklaşımlara göre; (1) Aşamalı Kümeleme Yöntemleri ve (2) Aşamalı Olmayan Kümeleme Yöntemleri biçiminde iki temel gruba ayrılmaktadır. Aşamalı Kümeleme Yöntemi olarak *Öklit uzaklığının* hesaplanmasını kullanan En Yakın Komşu yöntemi en basit yöntem olarak tek-bağ algoritmasında iki küme arasındaki uzaklık, her iki küme arasında yer alan kayıtlardan birbirlerine en yakın olanların uzaklığı olarak değerlendirilmektedir. Bu çalışmada da, bir kazı sonucu çıkartılan 58 adet toprak kaptaki element içerikleri En Yakın Komşu Yöntemi ile kümelendi ve kümeleme analizi SPSS 14.0 yazılımı ile gerçekleştirilmiştir. Bu çerçevede oluşturulan dendrogram incelendiğinde, tüm örnekleri içeren iki farklı ana kümeden söz etmek mümkündür. Bu iki kümedeki örneklerin iki ayrı bölgede üretildiği varsayımının geçerliliği kabul edilebilir.

Anahtar Kelimeler: Kümeleme tekniği, Kimyasal analiz, Orijin belirleme, SPSS, Dendrogram

Clustering Technics Using For Applied Sciences

Abstract: Main object of the clustering of the data is that the main heterogene data group are divided into different homogeneous data groups. Raw data has also a great importance for chemical analysis and the clustering analysis of these data bases on determining of the similarity or differences among them. Clustering methods are splitted up two groups; (1) Hierarchical Cluster Analysis Methods and (2) Nonhierarchical Cluster Analysis Methods. As to Nearest Neighbor method as Hierarchical Cluster Analysis Method the distance between two clusters is accepted that the nearest distances among the cluster data set for calculating. In this study the elements belongs to 58 ancient pots were analysed chemically and these data was subjected to clustering analysis by Nearest Neighbor method, and SPSS 14.0 was used as software for statistical evaluation. For that reason as examined prepared dendrogram it could be said there are two main clusters of the data set and as a result it was accepted that the ancient goods in these two clusters are produced at different places.

Keywords: Clustering analysis, Chemical analysis, Definition origin, SPSS, Dendrogram

1. Giriş

Günlük yaşamda ürünlerin pazarlanması ya da hizmetlerin sunulması gibi pratik alanlarda veya yapılan bilimsel ve teknolojik araştırmalarda toplanan verilerin değerlendirilmesinde karşılaşılan kümeleme tekniklerinin, gerek fizik, kimya ve biyoloji gibi temel bilimlerde gerekse tıp, mühendislik, nanoteknoloji, bilgi teknolojileri, genetik, çevre ya da biyoteknoloji gibi pek çok uygulamalı bilimde, madenlerin, ürünlerin, canlıların fiziksel ya da kimyasal özelliklerine göre orijinlerin belirlenmesi, yapılarının gruplanması, özelliklerinin zamana bağlı olarak değişimlerinin incelenmesi gibi alanlarda çok geniş kullanıma sahip olduğu pek çok kaynaktan açıkça

anlaşılmaktadır. Kimya alanında da benzer olarak, özellikle yapılan kemometrik analizlerde sağlanan ham veri artık hem çok miktarda hem de büyük öneme sahip olmaktadır, ve yapılacak olan bir deneysel çalışma sonrasında elde edilen verilerin hesaplamalarına ve değerlendirilmelerine geçmeden önce, özelliklerinin dikkatle incelenmesi, anlaşılır ve karşılaştırılabilir olması için istatistiksel olarak araştırılması, ve veri madenciliği ilkeleri çerçevesinde bütünleştirilmesi gerekmektedir [1].

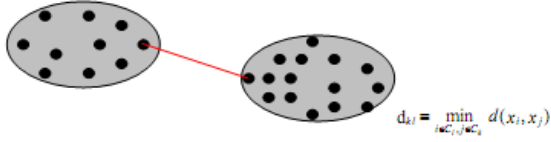
Kimyada deneysel verilerin değerlendirilmesi neden-sonuç ilişkilerinin belirlenmesi açısından istatistiğin ilkelerine, veri gruplarının oluşturulması ve anlamlandırılması açısından veri madenciliği

$$d_{ki} = \frac{d_{Ai} + d_{Bi}}{2}$$

Kümeleme yöntemleri, birim ya da değişkenleri uygun gruplara ayırırken grupları belirlemede izledikleri yaklaşımlara göre;

1. Aşamalı Kümeleme Yöntemleri (Hierarchical Cluster Analysis Methods)
 2. Aşamalı Olmayan Kümeleme Yöntemleri (Nonhierarchical Cluster Analysis Methods)
- biçiminde iki temel gruba ayrılmaktadır.

Aşamalı Kümeleme Yöntemi olarak *Öklit uzaklığının* hesaplanmasını kullanan Nearest Neighbor (En Yakın Komşu) yöntemi en basit yöntem olarak tek-bağ algoritmasında iki küme arasındaki uzaklık, her iki küme arasında yer alan kayıtlardan birbirlerine en yakın olanların uzaklığı olarak değerlendirilmektedir. Bu işlem Şekil 2’de ayrıntılı olarak gösterilmiştir:



Şekil 2. Tek-Bağ kümelemede iki küme arası uzaklık

Bu çalışmada, bir kazı sonucu çıkartılan 58 adet toprak kaptaki element içerikleri [5] En Yakın Komşu Yöntemi ile kümelendi. Kümeleme analizi SPSS 14.0 yazılımı ile gerçekleştirilmiştir.

2. Yöntem

Sunulan çalışmada; bir kazı bölgesinden çıkartılan 58 adet toprak kap örneklerinin kimyasal analizleri yapılmıştır. Element içerikleri kütlece % veya ppm düzeyinde olacak şekilde atomik absorpsiyon ile belirlenmiştir [5]. Çizelge 1 örnek bileşimlerine ilişkin verilerin saklandığı SPSS14 veri dosyası olup, yapılan kümeleme analizi sonuçları ve dendrogram ve Şekil 3’de verilmiştir.

Toprak kap örneklerinin kümeleme analizine ilişkin istatistiksel değerlendirme

Cluster [DataSet1] C:\Users\...\Kazi.sav

Case Processing Summary(a,b)

Cases		
Valid	Missing	Total

N	Percent	N	Percent	N	Percent
55	100.0	0	.0	55	100.0

a Squared Euclidean Distance used

b Single Linkage

Single Linkage (Çizelge 2’de verilmiştir)

Dendrogram (Şekil 3’de verilmiştir)

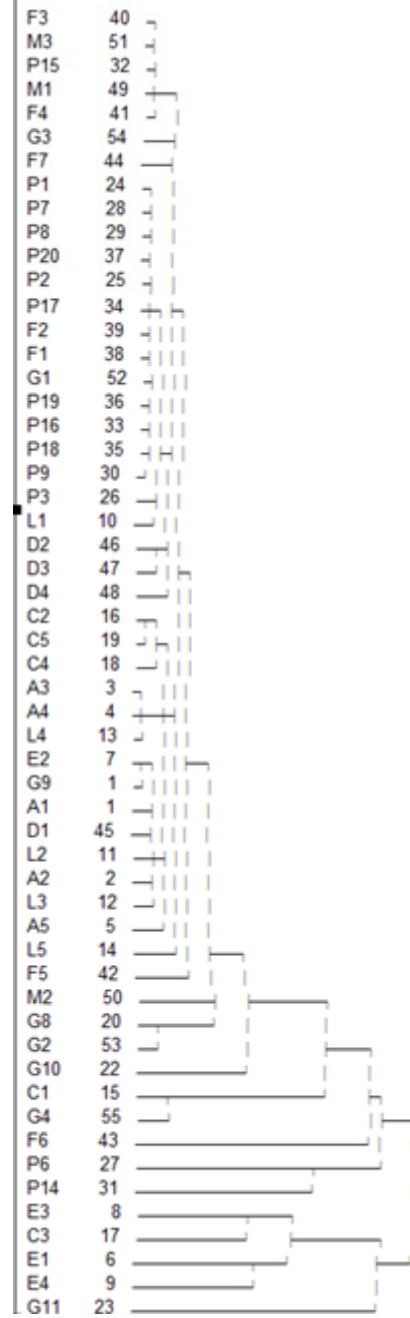
Çizelge 1. Toprak kap örneklerinin kimyasal analizleri

*SPSS_15_14_Data01_Dosya_Classify_HierarchicalClusterAnalysis1 [DataSet1] - SPS														
File Edit View Data Transform Analyze Graphs Utilities Add-ons Wi														
20 :														
	isim	örnek	Ti	Sr	Ba	Mn	Cr	Ca	Al	Fe	Mg	Na	K	sinif
1	A1	.1	.304	18	1007	642	80	1.840	8.342	3.542	.458	.548	1.799	A
2	A2	.2	.316	19	1246	792	84	2.017	8.592	3.696	.509	.537	1.816	A
3	A3	.3	.272	17	842	568	48	1.587	7.866	3.221	.540	.608	1.970	A
4	A4	.4	.301	14	843	526	62	1.032	8.547	3.455	.546	.664	1.908	A
5	A5	.5	.908	12	913	775	184	1.334	11.22	4.637	.395	.429	1.521	A
6	E1	.6	.394	10	1470	1377	90	1.370	10.33	4.543	.408	.411	2.025	A
7	E2	.7	.359	96	1188	839	86	1.398	9.537	4.099	.427	.482	1.929	A
8	E3	.8	.406	13	1485	1924	90	1.731	10.13	4.490	.502	.415	1.930	A
9	E4	.9	.418	13	1174	1325	91	1.432	10.50	4.641	.548	.500	2.001	A
10	L1	1.0	.360	11	410	652	70	1.129	9.802	4.280	.738	.476	2.019	A
11	L2	1.1	.280	11	1008	838	59	1.458	8.960	3.828	.535	.392	1.883	A
12	L3	1.2	.271	11	1171	681	61	1.456	8.183	3.285	.521	.509	1.970	A
13	L4	1.3	.288	10	915	568	60	1.268	8.465	3.437	.572	.479	1.893	A
14	L5	1.4	.253	10	833	415	193	1.228	7.207	3.102	.539	.577	1.972	A
15	C1	1.5	.303	13	601	1308	65	.907	8.401	3.743	.784	.704	2.473	A
16	C2	1.6	.264	12	878	921	69	1.164	7.926	3.431	.636	.523	2.032	A
17	C3	1.7	.264	11	1622	1674	63	.927	7.980	3.748	.549	.497	2.291	A
18	C4	1.8	.252	11	793	750	53	1.171	8.070	3.536	.599	.551	2.282	A
19	C5	1.9	.261	12	851	849	61	1.311	7.819	3.770	.668	.508	2.121	A
20	G8	2.0	.397	17	582	939	61	1.260	8.894	4.146	.656	.579	1.941	A
21	G9	2.1	.246	10	1121	795	53	1.332	8.744	3.689	.571	.477	1.803	A
22	G10	2.2	1.17	97	886	530	441	6.290	8.975	6.519	.323	.275	.762	A
23	G11	2.3	.428	45	1488	1138	85	1.625	9.822	4.367	.504	.420	2.055	A
24	F1	2.4	.259	38	399	443	176	11.60	5.901	3.283	1.37	.491	2.148	B
25	F2	2.5	.185	23	458	601	144	11.04	4.674	2.743	.711	.484	.909	B
26	F3	2.6	.312	27	383	682	138	8.430	6.550	3.680	1.15	.532	1.757	B
27	P6	2.7	.183	22	435	594	659	9.978	4.920	2.692	.672	.476	.902	B
28	P7	2.8	.271	39	427	410	125	12.00	5.997	3.245	1.37	.527	2.173	B
29	P8	2.9	.203	24	504	634	117	11.11	5.034	3.714	.728	.500	.864	B
30	P9	3.0	.182	21	474	520	92	12.92	4.673	2.330	.590	.547	.746	B
31	P14	3.1	.271	25	485	398	955	11.05	5.611	3.238	.737	.458	1.013	B
32	P15	3.2	.236	22	203	592	83	9.061	6.795	3.514	.750	.505	1.574	B
33	P16	3.3	.288	33	436	509	177	10.03	6.579	4.099	1.54	.442	2.400	B
34	P17	3.4	.331	30	460	530	97	9.952	6.287	3.344	1.12	.519	1.746	B
35	P18	3.5	.256	34	486	486	132	9.797	6.294	3.254	1.24	.641	1.918	B
36	P19	3.6	.292	28	426	531	143	8.372	6.874	3.360	1.05	.582	1.598	B
37	P20	3.7	.212	28	486	605	123	9.334	5.343	2.808	1.14	.595	1.647	B
38	F1	3.8	.301	32	475	556	142	8.619	6.914	3.597	1.08	.584	1.835	B
39	F2	3.9	.305	30	473	573	102	8.913	6.980	3.677	1.38	.616	2.077	B
40	F3	4.0	.300	20	192	575	79	7.422	7.663	3.476	1.06	.521	2.324	B
41	F4	4.1	.225	18	160	513	94	5.320	7.746	3.342	.841	.657	2.268	B
42	F5	4.2	.306	20	109	536	285	7.866	7.210	3.528	.971	.534	1.851	B
43	F6	4.3	.295	39	172	627	502	9.019	7.775	3.808	1.64	.766	2.123	B
44	F7	4.4	.279	23	99	760	129	5.344	7.781	3.505	1.20	.827	2.305	B
45	D1	4.5	.282	10	893	723	92	7.978	7.341	3.393	.630	.326	1.716	B
46	D2	4.6	.338	23	687	683	108	4.988	8.617	3.985	1.03	.697	2.215	B
47	D3	4.7	.327	15	686	590	70	4.782	7.504	3.569	.536	.411	1.490	B
48	D4	4.8	.233	98	580	678	73	8.936	5.831	2.748	.542	.282	1.248	B
49	M1	4.9	.242	18	182	647	92	5.303	8.164	4.141	.800	.734	1.905	B
50	M2	5.0	.271	47	198	459	89	10.20	6.547	3.035	1.15	.951	.828	B
51	M3	5.1	.207	18	205	587	87	6.473	7.634	3.497	.763	.729	1.744	B
52	G1	5.2	.271	19	472	587	104	5.119	7.657	3.949	.836	.671	1.845	B
53	G2	5.3	.303	23	522	870	130	4.610	8.937	4.195	1.08	.704	1.840	B
54	G3	5.4	.166	19	322	498	80	7.633	6.443	3.196	.743	.480	1.390	B
55	G4	5.5	.227	17	718	1384	87	3.491	7.833	3.971	.783	.707	1.949	B

Çizelge 2. Single Linkage

Stage	Cluster Combined		Coeffit	Stage Cluster First Appears		Next Stage
	Clus 1	Clus 2		Clus 1	Clus 2	
1	40	51	657.378	0	0	2
2	32	40	1006.104	0	1	11
3	29	37	1374.699	0	0	4
4	25	29	2090.121	0	3	7
5	34	39	2093.720	0	0	6
6	34	38	2217.303	5	0	7
7	25	34	3401.477	4	6	8
8	25	32	3542.388	7	0	9
9	25	36	3673.898	8	0	10
10	25	33	3680.334	9	0	14
11	32	49	4157.093	2	0	16
12	24	28	4382.173	0	0	20
13	3	4	4666.808	0	0	19
14	25	35	4798.097	10	0	15
15	25	30	5195.311	14	0	20
16	32	41	5636.538	11	0	34
17	16	19	6013.137	0	0	21
18	7	21	7614.867	0	0	25
19	3	13	8148.098	13	0	33
20	24	25	8868.688	12	15	23
21	16	18	13485.170	17	0	38
22	1	45	13751.280	0	0	24
23	24	26	13874.126	20	0	29
24	1	11	14648.362	22	0	25
25	1	7	14680.104	24	18	26
26	1	2	15662.465	25	0	27
27	1	12	15681.548	26	0	31
28	20	53	16269.501	0	0	44
29	10	24	16301.717	0	23	37
30	46	47	16465.311	0	0	35
31	1	5	18254.290	27	0	33
32	15	55	21477.332	0	0	50
33	1	3	21604.191	31	19	38
34	32	34	21698.034	16	0	36
35	46	48	22258.813	30	0	37
36	32	44	22963.883	34	0	39
37	10	46	23433.732	29	35	39
38	1	16	23971.363	33	21	40
39	10	32	24340.666	37	36	41
40	1	14	31608.972	38	0	41
41	1	10	31763.747	40	39	42
42	1	42	40402.016	41	0	43
43	1	30	55113.433	42	0	44
44	1	20	56443.633	43	28	45
45	1	22	77603.903	44	0	50
46	8	17	82629.018	0	0	48
47	6	9	91105.073	0	0	48
48	6	8	112097.485	47	46	53
49	27	31	128902.962	0	0	52
50	1	15	138654.697	45	32	51
51	1	43	170710.315	50	0	52
52	1	27	179095.838	51	49	54
53	6	23	181374.328	48	0	54
54	1	6	212813.939	52	53	0

kütleli homojen gruplara ayırmaktır. Kümeleme önceden tanımlanmış sınıfları esas alarak gruplamaz yani kümeleme gruplamada benzerlikleri kullanırken sınıflandırma önceden tanımlanmış sınıflar modelini temel alır.



Şekil 3. Toprak kap örneklerinin kümeleme analizine ilişkin çizilen dendrogram

3 Sonuçlar ve Tartışma

Veri tabanlarında farklı özelliklerdeki belli bir anlamı olmadan kayıtlı olan verilerin yorumlanarak anlamlı ve kullanışlı bilgiler haline getirilmesi veri madenciliğinin temelini oluşturmaktadır. Sınıflandırma ise veri madenciliğinin en çok kullanıldığı alandır ve sınıflandırmada yaklaşımlardan biri olan kümelemenin amacı ise heterojen olan ana

Bu çerçevede sunulan çalışmanın amacı da; bir kazı bölgesinden çıkartılan 58 adet toprak kap örneklerinin kimyasal analizlerinin yapılması ve elde edilen element içeriklerine göre kapların belirli kümeler altında toplanarak üretildikleri ve kullanıldıkları kazı bölgeleri hakkında ki düşünülenler ile gerçek değerler arasında bir ilişki kurmaktır. Bu çerçevede oluşturulan dendrogram incelendiğinde, tüm örnekleri içeren iki farklı ana bölgeden söz etmek mümkündür. Bu bölgeler E1, E3, E4, C3 ve G11 kodlu örnekler kümesi ile kalan diğer örnekler kümesinin kesinlikle aynı bölgede üretildiği anlaşılmaktadır. Ayrıca dendrogramın diğer alt birleşim kümeleri de incelenerek orijinlerle ilgili varsayımlara gidilebilir.

4. Kaynaklar

- [1] Odabaş, S., (2012). Objelerin benzer ya da farklı özelliklerine göre sınıflandırılmasında kümeleme tekniklerinin kullanılması, Yüksek Lisans, Muğla Üniversitesi, Fen Bilimleri Enstitüsü, Kimya Ana Bilim dalı.
- [2] Brereton R.G., (2003). Chemometrics data analysis for the laboratory and chemical plant, JohnWiley & Sons, Ltd.
- [3] Arifoğlu, U. (2005). *Matlab 7.04 Simulink ve Mühendislik Uygulamaları*, Alfa Ltd.
- [4] Özkan, Y., (2008). *Veri Madenciliği Yöntemleri*, Papatya Yayıncılık, Ankara.
- [5] Demir, C., Tokatlı, F., Ertaş, H., Özdemir, D., (2009). *Kemometri Yaz Okulu IIDers Notları*, I.Y.T.E. Fen Fak. Kim. Böl. İzmir.
- [6] Akpolat, O. (2010). *Matlab Uygulamaları ile Bilişim Teknolojileri*. Muğla Üniversitesi Yayınları, Muğla, 141-148s.
- [7] Han, J. ve Kamber, M., (2001). *Data Mining Concepts and Techniques.*”, Morgan Kauffmann Publishers Inc.
- [8] Karypis, G., Han,E.H.; Kumar,V. , (1999). CHAMELEON: A hierarchical clustering algorithm using dynamic modeling, *IEEE Computer*, 32(8).
- [9] Jain,A.K., Dubes,R.C., (1998). Algorithms For Clustering Data, *Prentice Hall*, Englewood Cliffs, New Jersey, 07632.
- [10] Kaya, H., Köymen, K., (2008). *Veri Madenciliği Kavramı ve Uygulama Alanları*, *Doğu Bölgeleri Araştırması*