HARRAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KÜMELEME ANALİZİ, KÜMELEME ANALİZİNE MATEMATİKSEL PROĞRAMLAMA YAKLAŞIMI VE BİR UYGULAMA

77775

İSMAİL YILDIZ

77775

DOKTORA TEZİ
ZOOTEKNİ ANABİLİM DALI
BİYOMETRİ VE GENETİK BİLİM DALI

ŞANLIURA -1998

HARRAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KÜMELEME ANALİZİ, KÜMELEME ANALİZİNE MATEMATİKSEL PROĞRAMLAMA YAKLAŞIMI VE BİR UYGULAMA

İSMAİL YILDIZ

DOKTORA TEZİ
ZOOTEKNİ ANABİLİM DALI
BİYOMETRİ VE GENETİK BİLİM DALI

Bu tez 04/09/1998 tarihinde aşağıdaki jüri tarafından değerlendirilerek oybirliği/~~oyçokluğu~~ ile kabul edilmiştir.

Doç.Dr.Hasan BAL	Doç.Dr.Yusuf YAYLI	Yrd.Doç.Dr.Rahmi KANAT
Jüri Başkanı	Üye	Üye(Danışman)

İÇİNDEKİLER

III. BÖLÜM

MATERYAL VE METOT
KÜMELEME ANALİZİNE MATEMATİKSEL PROĞRAMLAMA YAKLAŞIMI

ÖZET

DOKTORA TEZİ

KÜMELEME ANALİZİ, KÜMELEME ANALİZİNE MATEMATİKSEL PROĞRAMLAMA YAKLAŞIMI VE BİR UYGULAMA

İSMAİL YILDIZ

Harran Üniversitesi
Fen Bilimleri Enstitüsü Zootekni Anabilim Dalı
Biyometri ve Genetik Bilim Dalı
1998, Sayfa: 134

Bu çalışmada, öncelikle Kümeleme Analizi üzerinde duruldu. Kümeleme Analizine Matematiksel Proğramlama ile yaklaşan teknikler incelendi. Bir biyolojik uygulama yapılarak bunlar karşılaştırıldı.

Bu çalışma beş bölümden oluşmaktadır.

Birinci bölüm giriş bölümüdür.

İkinci bölümde genel bilgiler başlığı altında temel tanımlar ve tarihçe ele alınarak Kümeleme Analizi hakkında bilgi verilmiştir. Daha sonra Kümeleme Analizi tanımlanarak, Kümeleme Analizinin amacı, kümelemede kullanılan benzerlik ve uzaklık ölçütleri, kümeleme yöntemleri, küme sayısının belirlenmesi problemleri sunulmuştur. Ayrıca Diskriminant Analizi, Tercih Çözümü vb. kavramlara değinilmiştir.

Üçüncü Bölümde Materyal ve Metot açısından, Kümeleme Analizi detaylı bir şekilde verildikten sonra, Matematiksel programlama tekniklerinden olan Tamsayı Programlaması ve gruplandırma teorisi üzerinde duruldu. Daha sonra Tamsayılı proğramlamaya matematiksel formülasyonlar hazırlandı. Kümeleme Analizi

Metotlarının işini de yapan Matematiksel Proğramlama modelleri oluşturuldu. Bu Matematiksel modellere kümeleme analizinde yapılan işlemler uygulanıp kümeleme yapılmaya çalışıldı. Bu işlemlerden sonra Kümeleme Analizi ile yapılan gruplandırmalar ile Matematiksel Proğramlama formülasyonlarıyla yapılacak olan gruplandırmalar karşılaştırıldı ve bu iki yöntemin birbirine olan üstünlükleri tartışıldı.

Dördüncü Bölümde, yani uygulama bölümünde ise, Güneydoğu Anadolu Bölgesinin Diyarbakır İlindeki *Medicago* L. (Fabaceae) türü örnek verileri alınıp Kümeleme Analizi ve Kümeleme Analizinde Matematiksel Proğramlama kullanılarak kümeleme yapıldı. Bu kümeleme modelleri için bilgisayar proğramları da hazırlandı. Kümeleme Analizi (KA) ile yapılan kümelemeler ile Matematiksel Proğramlama (MP) ile yapılan kümelemeler, Akbayın ve Demir'in teşhis ettiği türler gözönüne alınarak karşılaştırıldı. Matematiksel Proğramlama modellerinin daha iyi kümeleme yaptığı gözlendi.

Beşinci ve altıncı bölümde ise, bulgular, tartışma ve sonuç ele alındı. Diskriminant Analizi ve Kümeleme Analizine alternatif olarak, verilerin kümelelenmesinde Matematiksel Proğramlama teknikleri kullanılabileceği gözlendi. MP tekniklerininin parametre dışı istatistiksel tekniklerinden biri olarak kabul edilebileceği ortaya konuldu.

ANAHTAR KELİMELER:
Diskriminant Analizi, Kümeleme Analizi, Lineer (Doğrusal) Proğramlama, Tamsayı Proğramlaması, Matematiksel Proğramlama.

ABSTRACT

DOCTOR OF PHILOSOPHY

CLUSTER ANALYSIS, A MATHEMATICAL PROGRAMMING APROACH FOR CLUSTER ANALYSIS AND APPLICATION

İSMAİL YILDIZ

Harran University
Graduate School of Natural and Applied Science
Department of Animal Science
Department Biyometry and Genetic
1998, Page: 134

In this study firstly, cluster analysis has been emphasized. The techniques aproaching mathematicall programming to the problems of clustering analysis have been studied, and they were compared by preparing a biological application.

This study consists of five chapters.

The first section is an introductory.

In the second chapter, the fundamental definition and the history of cluster analysis have been studied and introductory information has been provided. Then cluster analysis has been the defined and the objective of cluster analysis criteria used for similarity and the distance in clustering, clustering methods and the problems to determine the group numbers have been presented. In addition, the discriminant analysis, preference solution, etc. have been mentioned.

In the third section, the cluster analysis has been given in detail concerning material and method, integer programming and grouping theory which are the mathematical programming techniques focused on. Then, mathematical formulations for integer programming were preparaed. The mathematical programming models corresponding to cluster analysis method were formed. The operations performed in the cluster analysis were applied to these mathematical models to achieve grouping. Following these operations, the grouping performed in the cluster analysis and in the mathematical programming formulations were compared, and the advantages of these two methods have been discussed.

In the fourty chapter, which is the application chapter, the data on *Medicago* L. (Fabaceae) in Diyarbakır, a city in the Southeast Anatolia Region were groupped by mathematical programming for cluster analysis. Computer programs were prepared for the clusters models. The groups formed by the clustering analysis and these by mathematical programming were compared by considering the types suggested by Akbayın and Demir. The Mathematical programming models were proven to yield better grouping results.

In the fifth and sixth sections, the findings, discussions and the results were dealt with. It was concluded that mathematical programming techniques could be used as alternatives to the discriminant analysis and cluster analysis. The mathematical programming techniques could be taken as nonparametric statistical techniques.

KEYWORDS:
Discriminant Analysis, Cluster Analysis, Linear Programming, Integer Programming, Mathematical Programming

TEŞEKKÜR

Bu Doktora çalışmasını yöneten hocam sayın Yrd.Doç.Dr. Rahmi KANAT'a, çalışma süresince karşılaştığım güçlüklerde değerli yardımlarını esirgemeyen hocam sayın Doç.Dr. Hasan BAL'a, Doç.Dr. Salih ÇELEBİOĞLU'na ve veri te'mininde yardımlarını esirgemeyen sayın Doç.Dr. Hasan AKBAYIN'a teşekkür eder, saygılarımı sunarım.

Bu tez Harran Üniversitesi Araştırma Fon'unca desteklenmiştir(Proje no: 35). Desteklerinden dolayı Harran Üniversitesi Rektörü Sayın Prof.Dr. Mahmut SERT'e teşekkür eder, saygılar sunarım.

İsmail YILDIZ

ŞEKİLLER VE TABLOLAR

I. BÖLÜM

1.1. GİRİŞ

Kümeleme Analizi, konuyu yakından takip etmeyenler tarafından son derece sıkıcı, sıradan ve modası geçmiş bir kümeleme tekniği olarak kabul edilir. Kümeleme Analizi ile uğraşan kişinin bir ömür boyunca bir müzenin çekmeceleri içinde toplanmış tür örneklerini sınıflandırmaya çalıştığı düşünülebilir. Oysa Kümeleme Analizi, biyoloji araştırmalarında, yeryüzündeki milyonlarca canlı türünün belirli bir sistem içinde düzenlenmesi, Zooloji'de, Ziraat'te, özellikle Zootekni araştırmalarında, hayvanların ırklarına göre sınıflandırılması, en verimli hayvan yemi gruplarının oluşturulması, en iyi bal veren arı kovanlarının tespiti, Tıp'ta, hasta gruplarının tespiti, vb. araştırmalara uygulanabilir.

Kümeleme Analizi ile hiçbir ilgisi olmadığı düşünülen bir araştırmada bile biyolojik bir olay farklı canlı grupları arasında karşılaştırılıyorsa, Kümeleme Analizinin bu çalışmada önemli bir yeri vardır (2).

Kümeleme Analizi veya biyologların kullandığı Taksonomi, biyologların en önemli çabalarından biridir. Buna rağmen yeryüzündeki 5 milyon canlı türünden sadece 1,7 milyonu tanımlanıp sınıflandırılabilmiştir. Daha tanımlanıp sınıflandırılması gereken çok sayıda canlı türü olduğuna göre, Kümeleme Analizi biyologların en önemli çabalarından biri olmaya devam edecektir. Kümeleme Analizi, sınıflandırma işlemlerini sayısal temellere dayandıran bir bilim dalıdır (2).

Kümeleme Analizi çalışmalarında çok sayıda karakterden yararlanmak gerekmektedir. Kullanılan karakter sayısı artıkça, sınıflandırmanın daha iyi sonuçlar vermesi beklenir. Yapılan çalışmalarda kullanılan karakter sayısı 60 ve üzerinde olduğunda sınıflamanın oldukça kararlı olduğu gözlenmiştir (2).

Kümeleme Analizi ile uğraşanlar, sınıflandırma için biyokimyasal, kromozom davranış ve morfolojik karakterler kullanılmakla birlikte genellikle morfolojik karakterleri kolay ve ucuz elde ettikleri için onları kullanırlar.

Kümeleme Analizini her türlü sınıflandırma probleminin çözümü olarak görmek yanlış olur. Gelişen her bilim dalı gibi, Kümeleme Analizinde de bazı problemlerle karşılaşılabilir. Sayısal yaklaşımların Kümeleme Analizini Zootekni'de, hatta diğer bilim dallarında da daha etkin daha yararlı hale getireceği şüphesizdir.

Son zamanlarda Matematiksel Proğramlama birçok istatistiksel problemde kullanılmaya başlanmış ve bu kullanım her geçen gün yaygınlaşmaktadır. Bu çalışmada Matematiksel Proğramlama ile Kümeleme Analizi arasında ilişki kurulmak istenmiş, Kümeleme Analizi ile yapılan çalışmaların Matematiksel Proğramlama Teknikleriyle de yapılabileceği gösterilmeye çalışılmıştır. Bu nedenle önce Kümeleme Analizi üzerinde durulmuş, daha sonra Matematiksel Proğramlama tekniklerinden bahsedilmiştir. Kümeleme Analizi ile yapılan gruplandırmaların, Matematiksel Proğramlama ile daha kolay ve pratik olduğu gösterilmiştir. Ayrıca bir uygulama yapılarak, hem Kümeleme Analizi hem de Matematiksel Proğramlama kullanılmış ve bu iki tekniğin karşılaştırılması yapılarak biribirine olan üstünlükleri tartışılmıştır.

Kümeleme Analizinin genel amacı, gruplanmamış verileri benzerliklerine göre sınıflandırmak (gruplamak) ve araştırmacıya uygun, işe yarar özetleyici bilgiler elde etmede yardımcı olmaktır.

Bireylerin sınıflanması, ait oldukları kitlelerin (grupların) belirlenmesi ile uğraşan Kümeleme Analizinin amacı, sınıflama, nümerik taksonomi sözcükleri ile ifade edilmektedir. Kümeleme Analizi, çok değişkenli bir veri grubunda birbirine benzer veya yakın olan gözlemlerin bir arada gruplandırılmasını sağlayan bir yöntemdir. Çok yönlü amaçlar için kullanılabilir.

Kümeleme Analizinde çalışma yapmak isteyen bir araştırmacı, amacını çok iyi belirlemelidir. Değişken seçiminin sonucu etkileyen önemli bir faktör olduğu, yapılacak bir çalışmada rahatlıkla görülebilir. Değişken seçiminin hatalı olması, özellikle uzaklık türü çalışmalarda hatalı sonuçlar verir. Analizi birden fazla yöntemle sınamak, doğabilecek sonuçlardan araştırmacıyı uzaklaştırır.

Bireylerin gruplandırılmasında kullanılması nedeniyle Kümeleme ve Diskriminant Analizleri arasında benzerlik olmakla birlikte, iki yöntem arasında önemli farklılıklar bulunmaktadır. Diskriminant Analizinin bir bölümünde küme sayısı bilinmekte, bu sayı analiz süresince sabit kalmakta ve araştırmacıdan, bireyleri bu kümelere sınıflandırması istenmektedir. Kümeleme analizinde ise küme sayısı bilinmemektedir.

Kümeleme Analizi, değişkenler arasındaki uzaklık ölçüsüne dayanmaktadır. Kümeleme yöntemleri, hiyerarşik ve hiyerarşik olmayan yöntemler diye iki ana başlık altında toplanabilir. Bu yöntemler birbirine benzer ya da yakın gözlemleri araştırır ve yakınlık derecelerine göre gruplar oluşturur. Bireylerin gruplara atanması işlemi (Şehir Blok metriği, Öklid Metriği, Minkowski metriği, Tartılı Minkowski metriği, Tartılı Çebişev metriğ v.b. gibi) bazı uzaklık ölçütlerinden yaralanılarak yapılır. Bu çalışmada Öklid Metriği kullanılmıştır.

Kümeleme Analizi ile kümeler arası kareler toplamı ve kümeler içi kareler toplamı kriterleri ele alınmaktadır. P- boyutlu Öklid uzayında noktalar alındığı zaman, problemin lineer proğramlama problemi verilmiş olur. Bu problem yalnız kesin durumlar altında değer kazanır. Fakat avantaj olarak bir kümenin ayırma problemindeki iterasyon sayısını azaltır. Başlangıçta problemin lineer olmayan proğramlama formülasyonu verilir ve bu problemin özel durumda nasıl doğrusal proğramlama problemine indirgendiği araştırılır.

Yöntem olarak da kümeler arasındaki toplam uzaklığın küçültülmesi kriteri ele alınır. Son olarak kümeler arası en büyük uzaklığın küçültülmesine ilişkin bir

kriter verilir. Bu durumdaki formülasyon, *(0-1)* lineer tamsayılı proğramlama problemine eşdeğer olur.

Sonuç olarak Kümeleme Analizi ile, değerlendirilmek istenen değişkenler açısından birimler, gözlemlenen birimler açısından ise değişkenler kümelenir.

2. BÖLÜM

GENEL BİLGİLER

2.1. Temel Tanımlar ve Tarihçe

Literatürde Kümeleme Analizi ve Matematiksel Proğramlama ile ilgili farklı faktörler göz önüne alınıp yapılmış çalışmalar bulunmaktadır. Bu çalışmalarda doğa bilimleri, sağlık, ziraat, yerbilimi, mühendislik, politika ve ekonomide çeşitli modeller kurularak gruplandırmalar yapılmıştır. Özellikle Kümeleme Analizi ile ilgili son zamanlarda çok çalışma yapılmıştır. Öte yandan Matematiksel Programlama, Tamsayı Programlama ve Lineer Programlama konularında da birçok yeni çalışma yapılmıştır. Ancak Kümeleme Analizi ile Matematiksel Programlamayı birlikte ele alan çok az çalışma vardır.

Literatürde kümeleme işlemine kullanım yerlerine göre birçok araştırmada Q-Çözümlemesi (Q-Mode Analysis), Tipleme (Typology), Yığılma (Clumping), Sınıflandırma (Classification), Sayısal Sınıflama (Numerical Taxonomy), Örüntü Tanımlama (Pattern Recognition) ve Kümeleme Çözümlemesi(Analizi) (Cluster Analysis) gibi isimler verilmektedir.

Bu kümeleme işlemi gözlemler veya değişkenler için yapılmaktadır. Pek çok kaynakta değişkenler, karekter veya özellik terimleri ile eş anlamlı kullanılmaktadır. Gözlemler ise varlık veya gözlem birimi ile eş anlamlıdır.

Bazı araştırmacılar, Kümeleme Analizini değişkenler arası gruplandırma ve sınıflamayı da gözlemler arası gruplandırma olarak tanımlamışlardır (Kendall ve Stuart, (1966)). Fakat genellikle Kümeleme Analizinin anlatıldığı kaynaklarda kümeleme, gözlemler arası gruplandırma olarak tanımlanmaktadır.

Kendall ve Buckland (1960) küme tanımını; " Küme, İstatistiksel verinin yakın öğelerinden oluşan gruptur " şeklinde vermişlerdir (Kendall ve Buckland, (1960)).

Verilerin kümelendirilmesinde ilk kullanım biyoloji ve zoolojide olmuştur. Hindistan'da eski Roma ve Yunan medeniyetlerinde, insanlar çeşitli özellikleri bakımından gruplandırılmışlardır. Bu bakımdan ilk basılı eser 18. yüzyılda İngiltere 'de yaşıyan bitki ve hayvanların sınıflandırılması için verilmiştir. Doğa bilimlerindeki asıl gelişme 1930'lu yıllarda olmuştur (Everitt, (1974)). Bu yıllarda Tyron ilk defa "Küme" terimini faktör çözümlemesine ve temel bileşenler çözümlemesine alternatif olarak ortaya atmıştır (Duran ve Odell, (1971)).

Kümeleme Analizinde 1960 'dan sonraki dönem, önemli gelişmelerin olduğu dönemdir. 1960-1970 yılları arasındaki kaynaklardan çok sayıda kümeleme yönteminin bu dönemde ortaya çıktığı görülmektedir. Bu dönemin son yıllarında Matematikçiler ve Matematiksel İstatistikçiler, kümeleme analizine daha uygun bir yaklaşım yapma ihtiyacını duymuşlardır. Bunun önemli sebeplerinden birisi, bu döneme kadar önerilen metotların özelliklerini inceleyen bir matematiksel çerçeve kullanılmamış olmasıdır.

Yine bu dönemde, değişik metotları aynı verilere uygulayan karşılaştırmalı çalışmaların arttığını görmekteyiz. Gruplandırma problemleri için son zamanlarda bir çok Matematiksel Proğramlama modelleri alternatif olarak sunulmuştur. Bunlardan bazıları ; Rao, (1971); Freed ve Glover ,(1981); Freed ve Glover , (1986); Bruce, (1991); Lam ve Moy, (1996); Lam ve Moy, (1997)'dır.

Kümeleme Analizi problemini çözmek için pek çok geleneksel metot ileri sürülmüştür. Tek Bağlantı, Tam Bağlantı, Grup Ortalama Bağlantı, Küresel Ortalama Bağlantı, Ortanca Bağlantı, Esnek Kümeleme ve En Küçük Varyans Kümeleme (Ward Metodu) yöntemleri aşağıdan yukarıya doğru hiyerarşik metotların bazı önemlileri olup, Kümeleme Analizinde kullanılmaktadır (Blashfield and Aldenderfer, (1978); Gordon, (1981); Johnson and Wichern, (1988)). Tepe tırmanma ve k-ortalama metotları (Macqueen, (1967)) iki iteratif yeniden atama posedürleridir.

Ama geleneksel yaklaşımlar herhangi bir kritere göre genel bir optimaliteyi garantilememektedir. Geleneksel metotların yetersizliği, Kümeleme Analizine yönelik Matematiksel Proğramlama yaklaşımlarının incelenmesini ciddi olarak gözönüne alınmasına sebep olmuştur (Vinod, (1969); Rao, (1971); arthanari and Dodge (1981); Aronson and Klein, (1989)). Matematiksel Proğramlama Metotlarının avantajı, spesifik optimizasyon amaçlarına yönelik optimal çözümler sağlamasıdır. Kümeleme için uygun bir kriterin mevcut olduğu durumlarda bu en faydalısıdır. Mesela, Matematiksel Proğramlama yaklaşımlarının pek çok yer problemine başarı ile uygulanmaktadır.

Kümeleme Analizini içeren çok sayıda çalışma bulunmaktadır. Ancak Kümeleme Analizi ile Matematiksel Proğramlamayı beraber inceleyen çok az çalışma vardır;

Çetinel (1982), "Çok değişkenli verilerin kümelendirilmesi için istatistiksel bir yöntem" çalışmasıyla bu konuya yaklaşan istatistiksel metotlar ortaya koymuş. Lam(1991), "Linear Goal Programming in Classification and Preference Decomposition" çalışmasıyla konuyu derinlemesine incelemeye çalışmış ve bazı matematiksel modeller geliştirmiş. Nadar (1993), "An approach to optimization method of cluster analysis" çalışmasıyla bazı optimizasyon tekniklerini incelemeye çalışmıştır. Biz çalışmamızda, Lam (1991)'ın öne sürdüğü modellere dayanarak Kümeleme Analizini inceleyip, Kümeleme Analizine bazı Matematiksel Proğramlama teknikleriyle yaklaşılıp, farklı metrik kullanılarak bu teknikler *Medicago* L.türlerine uygulanmıştır.

Çoğu zaman bir çok nesnenin, kişinin, değişkenin ve sembolün az sayıda birbirinden tamamen bağımsız kümelere gruplandırılması gerekir. Böylece bir grup içindeki üyeler bazı yönlerden diğerlerine benzerdir. Böyle yapılan bir kümeleme genellikle belli bir bilgi kaybına veya ölçülebilen bir değere sebep olabilir. Bu kümeleme problemine sosyal bilimlerde olduğu gibi çeşitli tabii bilimler literatüründe de biraz önem verilmiştir.

Kümeleme için kullanılmakta olan tekniklerde, tamsayı proğramlaması esasına dayanan bir görüş noktası yoktur. Ward ve yardımcıları (22), Fisher (23)'in "aşamalı yaklaştırma tekniği" 'ne benzeyen bir "hiyerarşik kümeleme yöntemi" geliştirmişlerdir. Yeteri kadar küçük bir m küme sayısı elde edene kadar, küme sayısını n'den n-1'e kadar her iterasyonda azaltılır.

Ward ve Fisher tarafından savunulan yöntemler, tam sayımın tüm aşamalarını n'den m'ye indirgeme ve bu indirgemenin her aşamasında alternatif kümeleme metotları ile ilgilidir.

Bu algoritmalar dallandırma ve sınıflandırma teknikleri kullanılarak geliştirilebilir. Kümeleme probleminin bir tamsayı proğramlaması formulasyonu oldukça kullanışlıdır. Çünkü bu tamsayımı gerektirmemektedir. Ball ve Hall (24) tarafından geliştirilen bir iteratif teknikte Matematiksel Proğramlamada alınan binlerce benzer adımlar vardır. Ball ve Hall "tipik" küme noktalarının keyfi seçimini yaparak başlarlar. Matematiksel Proğramlamada bir esas çözüm ile başlanır ve en uygun çözüme G.B.Dantzig ve diğer arkadaşları tarafından ispatlanmış, bazı çok iyi bilinen teoremler ile de garanti edilmiştir. En iyi kümelemenin elde edilmesine yönelik iteratif prosedürü yakınsama sorunu, Ball ve Hall (24) tarafından gözardı edilmiştir.

Gruplandırma problemine yönelik yeterli bir çözümün elde edilmesi amacıyla doğrusal proğramlama çözümleri düzlemlerin kesilmesi, dal-sınır ve mağaza yeri problemleri ile ilişkili tamsayı proğramlamasının kullanımı için bir araç geliştirmek mümkün olmalıdır.

Çok değişkenli analizlerde lineer proğramlama yaklaşımları pek çok araştırmacı tarafından incelenmiştir.

Sınıflama, nesnelerin uygun sınıflara atanması problemi ile ilgilenir. Sınıflama yaklaşımının iki ana türü Kümeleme Analizi ve Diskriminant Analizi'dir.

Kümeleme Analizi "benzer" nesnelerin başlangıçta tanımlanmamış olan sınıflara göre "gruplanması" ile ilgilenir. Kümeleme Analizi yapmanın en önemli amacı verilerin basitleştirilmesidir. Diskriminant Analizi ise, nesnelerin bilinen bazı popülasyonlara "ayrılması" ile ilgilenir. Diskriminant Analizinin gerçekleştirilmesinin ilk amacı, yeni nesneleri doğru popülasyona atamaktır. Geleneksel yaklaşımlar ve İstatistiksel yaklaşımlar dahil pek çok metot, sınıflama probleminin iki türünün de çözümlenmesinde kullanılabilir.

Kümeleme Analizi'nde, geleneksel yaklaşımlar herhangi bir kritere göre optimal çözümler garantilemez. Sonuç olarak, Kümeleme Analizine yönelik değişik Matematiksel Proğramlama modelleri geliştirilmiştir. Matematiksel Proğramlama modellerinin avantajı, spesifik optimizasyon amaçlarına yönelik optimal çözümler sağlamasıdır. Biz Kümeleme Analizinde Lam (10)(Lam, Kim Fung Bruce) tarafından geliştirilen modellerde kullanılan Şehir Blok metriği yerine Oklid metriği kullanarak bunları inceleyeceğiz. Akbayın ve Demir (1993) tarafından elde edilen veri kümesini, Lam (10) tarafından geliştirilen modellerden ve diğer İstatistiksel Kümeleme metotlarından elde edilen küme çözümlerinin performans özelliklerini incelemek için kullanacağız. Küme uzaklıklarının yedi farklı kriteri ve küme uzaklıklıkları arasındaki bir kriter ile birlikte dokuz kriter, Uygulama bölümünde kullanılan biyolojik verilerden elde edilen küme çözümlerini değerlendirmek üzere kullanılmıştır. Lam (Lam, Kim Fung Bruce) modellerinin tamamı geleneksel yaklaşımlara göre çok daha iyi küme çözümleri sağlamaktadır (10).

Özetle, bu çalışmada Kümeleme Analizi, Diskriminant Analizi ve tercih çözümündeki problemleri çözmek için, bazı doğrusal amaç proğramlama modelleri incelenmiştir. Matematiksel Proğramlama ile oluşturulan modeller, Kümeleme Analizindeki modellere göre ilave optimizasyon hedefleri sağlamaktadır. Kümeleme Analizindeki alternatif optimal amaçların seçimindeki esnekliğin artışı, Kümeleme Analizine yönelik Matematiksel Proğramlama yaklaşımlarının uygulama sayısının daha fazla artmasını motive edecektir.

Hedef Proğramlama modeli ilk kez Charnes, Cooper ve Ferguson (26) tarafından 1955 yılında önerildi. Onların çalışması, çoklu regresyon içeren çok değişkenli analizler, Kümeleme Analizleri, Diskriminant Analizi ve tercih seçimindeki Doğrusal proğramlama tekniklerini kullanarak, pek çok uygulama yapılmasına sebep oldu (Wagner, (1959); Vinod, (1969); Rao, (1971); Srinivasan ve Shocker, (1973); Freed ve Glover, (1981a)). Tercih çözümü, bir ferdin çok özellikli tercih yapısını, o ferdin relatif tercihlerini kullanarak ölçüm yapan metotların bir sınıfıdır.

2.2. KÜMELEME ANALİZİ

Kümeleme Analizi iş, endüstri, tıp, ziraat ve eğitim gibi bilimsel araştırmanın pek çok alanında uygulandığını belirtmiştik. Kümeleme problemlerini çözmek için pek çok geleneksel metot sunulmaktadır. Fakat bu metotların başarılı olamadığı kümeleme problemleri bulunmaktadır. Bu nedenle Kümeleme Analizine yönelik Matematiksel Proğramlama modelleri geliştirilmiştir. Kümeleme Analizi için Matematiksel Proğramlama 'nın amaç fonksiyonları genellikle ya "küme uzaklıkları içerisinde" veya "küme uzaklıkları arasında"'ki ölçümler olarak tanımlanmaktadır. Çok kriterli optimizasyon çerçevesinin bir benzeri ile biz Kümeleme Analizi için anlamlı kriterlerin bir bütün dizisini oluşturmak için sistematik bir yol sağlıyoruz. Aynı zamanda yedi Matematiksel Proğramlama modeli sunmakta ve 2. bölümde mevcut iki modeli de yeniden formüle etmekteyiz. Ayrıca hazır bir veri kümesi kullanılarak bizim modellerimiz ve diğer bazı popüler kümeleme prosedürlerinin performans özellikleri incelenecektir.

Küme, bir veri grubunun birbirine komşu olan nesnelerinin meydana getirdiği bir topluluk olarak tanımlanabilir. Başka bir tanımlamada da küme, istatistiksel kitlenin yakın elemanlarının bir grubu olarak verilmektedir (Kendall ve Buckland, (1960)). Çok geniş anlamıyla kümeleme, bazı hususlarda herbiri diğerine benzer olduğu düşünülen nesnelerin bir arada toplanması işlemidir.

1960-1970 yılları arasında, değişik metotları aynı verilere uygulayan karşılaştırmalı çalışmaların arttığını görmekteyiz. Bu tür çalışmalara örnek olarak, varyans çözümlemesinin temel bağıntısı olan,

T : Genel Kareler Toplamı

W : Gruplar İçi Kareler Toplamı

B : Gruplar Arası Kareler Toplamı

olmak üzere;

$$T = W + B \qquad (2.1)$$

denkleminden hareket ederek, en iyi küme yapısını belirleyen ölçütlerin kullanılmasını gösterebiliriz (Edwards ve Cavalli- Sforza, (1965); Friedman ve Rubin, (1967)).

Bu dönemin ilk yarısından başlayarak günümüze kadar yapılan çalışmalarda belirli ölçütleri enbüyükleme veya enküçükleme ilkesine bağlı kümeleme yöntemlerinin de kullanıldığını görmekteyiz. Bu çalışmaların amaçları, ya verilen herhangi bir kümeleme probleminde küme sayısını belirlemek veya kümelerin sayısı için yaklaşık güven aralıkları oluşturmaktır.

Kümeleme Analizinin kullanılabilir sonuç vermesi, bu analizin çok değişkenli çözümleme yönteminin uygulandığı hemen hemen tüm bilim dallarındaki kaynaklarda rastlanmasına neden olmuştur. Dolayısıyla doğa bilimleri başta olmak üzere, sağlık, ziraat, yerbilimi, mühendislik, politika ve ekonomide ortaya çıkan tüm çok değişkenli veriler kümeleme analizine konu olmuştur.

2.2.1. Kümeleme Analizinde Sayısal Yöntemler

Çalışmamızda nesne terimiyle sınıflandırılan verilerin herbiri, karakter terimiyle de nesnelerin ölçülen morfolojik, biyokimyasal, davranış gibi özelliklerine ilişkin birer değişken anlaşılmalıdır. Metrik veriler için özellikle uygun olan parametrik model, nesnelerin uzayda temsil edilmeleri esasına dayalıdır. Örneğin, eğer bir nesne için iki değişken ölçülmüş ise, bu durumda uzay iki boyutludur ve durum basit bir X-Y grafiği ile gösterilebilir. Bu durumda birbirine dik açılardaki eksenlerden herbiri bir değişkene karşılık gelir ve nesnenin düzlemdeki yeri de bu eksenler üzerindeki değerlere bağlı olarak belirlenir. Bu modeli üçüncü bir değişken ilave ederek üç boyutlu hale getirmek kolaydır. Fakat dört, beş ve daha fazla boyutlu değişkenleri, yani eksenleri fiziksel olarak algılamak oldukça zordur. Yine de bir nokta, sonsuz sayıda eksene (boyuta) dayanarak tanımlanabilir. Boyutların değişkenleri temsil ettiği bu modellere çok boyutlu model denir. Nesneler çok boyutlu uzayda belirlendikten sonra Kümeleme Analizi ile uğraşan kişinin hedefi bu noktaların dağılımındaki hiyerarşik yapıyı ortaya çıkarmaktır (2)

2.2.2. Kümeleme Analizinin Amacı

Kümeleme yönteminin amacı, kullanıcının amacına bağlı olmakla birlikte, genel olarak aşağıdaki şekilde verilmektedir:

a- Uygun modeli bulmak,

b- Model uyumu sağlamak,

c- Gruplar içi ön tahmin hesapları yapmak,

d- Ön tahmin sınaması yapmak,

e- Veri yapısını bulmak,

f- Ön tahmini genelleştirmek,

g- Boyut indirgemek

'tir (Everitt, (1974)).

Uygulamada fazla sayıdaki çok değişkenli gözlemleri tanımlamak oldukça problemli olmaktadır. Bu amaçla istenilen bilgiyi ortaya koymak ve özetlemek gerekir. Bu nedenle, araştırmacı gözlemleri kontrol edilebilir gruplara ayırmak zorundadır. Aksi halde bu gözlemleri her yönüyle kontrol edemez. Bu nedenle kümeleme analizi "veri düzenleme" işlemini yapmak için de kullanılabilir. Belirlenen bir ölçüye göre gözlemler gruplara ayrıldığında, gruplar içinde homojenlik ve aynı zamanda gruplar arasında heterojenlik sağlanmalıdır. Bu takdirde araştırmacı, çok sayıda gözlem yerine tanımlayıcı birkaç küme üzerinde çalışma kolaylığına kavuşur. Bu durum araştırmacıya zaman ve hesaplama gibi birçok yönde avantaj sağlar. Ancak yapılan bu işlemle bir miktar bilgi kaybı söz konusu olacağından, Kümeleme Analizinin amacı bu bilgi kaybını en düşük seviyede tutmaktır. İlgili bilgi kaybını ölçen bir kayıp fonksiyonu dikkate alındığında, en iyi kümeleme teorik kayıp fonksiyonunun en küçüklenmesi ile bulunabilmektedir.

Kümelemenin bir diğer amacı, gözlemlerdeki doğal yapıyı belirlemektir. Bu bakımdan, Kümeleme Analizi bu doğal durumu ilgilendiren ön tahminleri ortaya çıkarmada kullanılabilir. Kümeleme Analizinde en iyi yararı sağlayabilmek için aşağıdaki hususlara dikkat edilmelidir:

a- Kümeleme yöntemleri ön tahmin sınaması için bir araçtır.

b- Ortaya çıkan kümeler kesin sonuç olmayıp, olası görünümler olabilir.

c- Kümeleme Analizi sonunda, veri kümesinin karışık yapısı ortaya çıkabilir.

d- Veri yapısının gerçek görünümü ortaya çıkabilir.

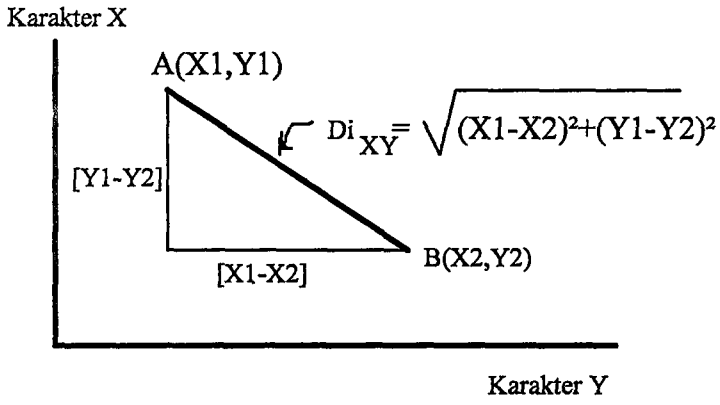
e- Verilerde küme yapısının olmaması ya da yalnızca bir küme olması gibi iki olası durumla karşılaşılabilir (Anderberg, (1973)).

Birbirine benzer nesneleri aynı kümeye, benzemeyenleri de farklı kümelere yerleştiren bu yöntemde, önce noktaların yani nesnelerin birbirlerine uzaklıkları hesaplanır ve bu uzaklıklara dayanarak kümeler genellikle hiyerarşik bir düzende oluşturulurlar. Kümeleme Analizinin bütünleyici algoritmasına göre kümeleme işleminin aşamaları aşağıdaki şekilde görülmektedir.

Şekil 1. : Bütünleyici algoritmaya göre kümeleme işlemi (2)

2.2.3. Kümelemede Kullanılan Benzerlik ve Uzaklık Ölçütleri

Bir düzlemde yer alan iki nokta arasındaki uzaklığı bulmada bir dik üçgenin hipotenüsünün, diğer iki kenar sayesinde hesaplanması esasına dayanan Öklid uzaklığı kullanılabilir. Bu uzaklık ölçüsü üç veya daha fazla karakter boyutu için de geçerlidir.



Şekil 2.: İki boyutlu uzayda A ve B nesnelerinin arasındaki Öklid uzaklığı ($D_{i_{xy}}$) ve Manhattan uzaklığı $d(x,y) = |x_1 - x_2| + |y_1 - y_2|$ şeklindedir