ÖZNİTELİK BÖLÜMLEMESİ VE ÇAKIŞAN ÖZNİTELİK ARALIKLARI ŞEKLİNDE ÇOKLU KAVRAM TANIMLARI İÇİN TOPLU MAKİNA ÖĞRENMESİ TEKNİKLERİNİN GELİŞTİRİLMESİ

PROJE NO: EEEAG-153

H. ALTAY GÜVENİR

KASIM 1997
ANKARA

ÖNSÖZ

“Öznitelik Bölümlemesi ve Çakışan Öznitelik Aralıkları Şeklinde Çoklu Kavram Tanimlari İçin Toplu Makina Öğrenmesi Tekniklerinin Geliştirilmesi” adlı ve EEEAG-153 numaralı bu proje TUBİTAK’ın desteğiyle Bilkent Üniversitesi Bilgisayar ve Enformatik Mühendisliği Bölümünde ve Doç. Dr. H. Altay Güvenir’in yürütücüğünde gerçekleştirilmiştir. Proje 1 Ağustos 1995 de başlamış ve 1 Ağustos 1997 de tamamlanmıştır. Projede Aynur Akkuş ve Gülşen Demiröz adlarında iki araştırma görevlisi de görev almıştır.

İçindekiler

Tablo Listesi **3**

Şekil Listesi **4**

Öz **5**

Abstract **6**

1 GİRİŞ **7**

2 DAHA ÖNCEKİ ÇALIŞMALAR **12**

3 DERLENEN VERİKÜMELERİ **14**

Dermatoloji Verikümesi 14
Aritmi Verikümesi 16

4 ÖZNİTELİK BÖLÜMLEMESİ ŞEKLİNDE ÇOKLU KAVRAM TANIMLARININ ÖĞRENİLMESİ **18**

5 ÇAKIŞAN ÖZNİTELİK ARALIKLARI ŞEKLİNDE ÇOKLU KAVRAM TANIMLARININ ÖĞRENİLMESİ **26**

6 ÖZNİTELİK AĞIRLIKLARININ GENETİK ALGORİTMA İLE ÖĞRENİLMESİ **31**

7 SONUÇ **37**

Kaynakça **39**

Ekler **42**

Tablo Listesi

Tablo 1. Dermatoloji verikümesindeki sınıflar ve örnek sayıları. 15

Tablo 2. Dermatoloji verikümesindeki öznitelikler. 15

Tablo 3. Aritmi verikümesindeki sınıflar ve örnek sayıları. 17

Tablo 4. BCFP.0, BCFP.1, BCFP.2, BCFP.3 ve BCFP.4 programlarının bazı doğal verikümeleri için elde edilen 5-katlı çapraz sağlama yolu ile elde edilen sınıflandırma doğrulukları (% olarak). 25

Tablo 5. BCOFI Algoritmasının bazı doğal verikümeleri için elde edilen 5-katlı çapraz-sağlama sınıflandırma doğruluğu ve programın Sparc 20/61 iş istasyonundaki ortalama çalışma süreleri. Bir karşılaştırma olması için aynı değerler 1-NN için de verilmiştir. 27

Tablo 6. BCFP.4 ve BCOFI Algoritmalarının bazı doğal verikümeleri için elde edilen 5-katlı çapraz-sağlama sınıflandırma doğrulukları. Burada her bir sınıflandrıcı için önce öznitelik ağırlıklarının eşit alındığı durumdaki doğrulukları sonrada genetik algoritma ile öğrenilen ağırlıklar kullanılarak elde edilen doğruluklar verilmiştir. Ayrıca genetik algoritmanın bir adım ilerlemesi için SUN-Sparc 20/61 iş istasyonunda gereken süre verilmiştir. 35

Şekil Listesi

Şekil 1. Tipik bir EKG işareti. 17
Şekil 2. BCFP Algoritmalarının genel yapısı. 19
Şekil 3. BCFP.1 algoritmasıyla oluşturulan örnek öznitelik bölüntüleri. 20
Şekil 4. BCFP.1 algoritmasının sınıflandırma yordamı. 21
Şekil 5. BCF1 algoritmasıyla örnek bir sınıflandırma. 22
Şekil 6. BCFP.1 algoritmasında kullanılan bölüntüAra yordamı. 22
Şekil 7. BCFP.2 algoritmasında kullanılan bölüntüAra yordamı. 23
Şekil 8. BCFP.3 algoritmasının sınıflandırma yordamı. 24
Şekil 9. BCFP.3 algoritmasının bölüntü_ağırlığı yordamı. 24
Şekil 10. İki öznitelikli bir verikümesi ve çakışan öznitelik aralıkları şeklinde gösterimi. 26
Şekil 11. BCFI Algoritması. 28
Şekil 12. BCOFI tarafından dermatoloji verikümesi için üretilen çıktının bir bölümü. 29
Şekil 13. Genetik Algoritma. 33
Şekil 14. Bir kromozonun uygunluk değerinin belirlenmesi. 34

Öz

Sınıflandırma problemleri çoklu kavram tanımlaması problemleri şeklinde modellenebilmektedir. Çoklu kavram tanımlarının sınıflandırılmış örneklerden öğrenilmesi makina öğrenmesi alanında yapılan çalışmalar arasında en çok uygulama alanı bulmuş olanlardan birisidir. Bu projede kavram tanımlarının hem öznitelik bölümlemesi ve hem de çakışan öznitelik aralıkları şeklinde toplu şekilde öğrenilmesi için algoritmalar geliştirildi. Bu çalışmada kavramlar her özniteliğin bir bölümlemesi veya her birisi bir kavrama karşı gelen çakışan öznitelik aralıkları şeklinde tanımlanmaktadır. Verilen yeni bir vakanın sınıflandırılması ise ayrı ayrı özniteliklere göre gerçekleştirilen sınıflandırılmalar arasında yapılan bir ağırlıklı oylama sonucunda belirlenmektedir. Özniteliklerin ağırlıklarının da öğrenilmesi sayesinde ilgili öznitelikler ilgisiz olanlardan ayrılabilmekte, bunun da ötesinde özniteliklerin bağıl önem dereceleri de belirlenebilmektedir. Bu projede öznitelik ağırlıklarının öğrenilmesinde genetik algoritmalardan yararlanılması da incelenmiştir. Bu amaçla, yeni bir çaprazlama operatorü de geliştirilmiştir.

Proje sırasında geliştirilen teknikler sayesinde çeşitli alanlarda çoklu kavram tanımlamalarının mekanik olarak öğrenilebilmesi ve her özniteliğin diğer özniteliklere göre önem derecesinin belirlenebilmesi mümkündür. Geliştirilen öğrenme tekniklerinin gerçek hayat problemlerine uygulanabilirliğinin sınanmasına yönelik olarak, biri EKG işaretlerinden aritmi karakteristiklerinin tanınması, diğeri de erythemato-squamous hastalıklarının histopatolojik tanımı için olmak üzere iki tıbbi verikümesi oluşturulmuştur. Proje kapsamında geliştirilen teknikler hem bu verikumeleri ve hemde UCI-Repository'deki standart verikumeleri üzerinde sınanmış ve çok başarılı sonuçlar alınmıştır.

Anahtar kelimeler: Makina öğrenmesi, kavram öğrenme, genetik algoritmalar, sınıflandırma.

Abstract

Classification problems can be modelled in terms of multi-concept description problems. Induction of multi-concept descriptions from classified examples is one of the fields machine learning research that have found large number of applications to real-world problems. In this project, we have developed non-incremental algorithms for learning concept descriptions in the form of feature partitions and overlapping feature intervals. Here, concepts are represented as the partitioning of the values for each feature or the set of overlapping feature intervals corresponding to concepts. The classification of a novel instance is determined through a weighted voting scheme on classifications based on the individual features. Because the weights of features can also be learned from the training dataset, it is possible to identify the relevant features from the irrelevant ones; moreover, it is possible to learn the relative importance of the features for classification. In this project, the use of genetic algorithms for learning the feature weights has been investigated. For this purpose, a new crossover operator has been developed.

Using the techniques developed during the project, it is possible to mechanically learn multi-concept descriptions and feature weights in large number of domains. In order for the evaluation of the learning algorithms developed, we have compiled two medical datasets; one for the description of arrhythmia characteristics from ECG signals, and the other for the differential diagnosis of erythemato-squamous diseases. The learning techniques developed in the project have been tested on these datasets along with the standard datasets from the UCI-Repository, and very successful results were obtained.

Key words: Machine learning, concept learning, genetic algorithms, classification

1 GİRİŞ

Bilindiği gibi zekanın en önemli öğelerinden birisi öğrenme yeteneğidir. İnsanlar, hatta bazı hayvanlar, geçmiş deneyimlerinden dersler almakta, bu öğrendikleri sayesinde, benzer durumlarla karşılaştıklarında daha doğru kararlar verebilmektedirler.

Bilgisayar bilimleri arasında, son yıllarda hız kazanan araştırma alanlarından birisi de *makina öğrenmesi* üzerinde yapılan çalışmalardır (Mitchell, 1997). Burada amaç bilgisayarların da, aynı insanlarda ve diğer canlı varlıklarda olduğu gibi, öğrenebilmelerini sağlamaktır. Bu tür programlar öğrenilecek konunun tam bir matematiksel modelinin bulunmadığı durumlarda yararlı olabilmektedirler. Tıp ve eknominin birçok dalı da işte bu nedenle makina öğrenmesinin uygulama alanına girmektedir.

Tüme varım tabanlı makina öğrenmesi algoritmaları *öğrenme* ve *karar verme* adı verilen iki aşamadan oluşurlar. Öğrenme aşaması, genellikle, geçmiş deneyimleri içeren örneklerden yola çıkıp, çeşitli tüme varım yöntemlerini kullanarak öğrenilen konunun bir modelini çıkarır. Karar verme aşaması ise bu modeli kullanarak, yeni ortaya çıkan durumlar hakkında doğru kararlar vermeye çalışır (Langley, 1996).

Öğrenen bir bilgiyar programından beklenen en önemli özellik öğrendiği alanda doğru kararlar verebilmesidir. Bunun yanında öğrenme ve karar verme sürelerinin kısalığı ve defolu (bir miktar hatalı veya eksik bilgi içeren) örneklerden de öğrenebilme de önemlidir. Ayrıca, öğrenmede kullanılan örnekler konu ile doğrudan ilgili bilgiler dışında, tamamen veya daha az ilgili bilgiler de içerebilmektedir. İyi bir öğrenme programı ilgili bilgileri ilgisizlerden ayırabilmeli, ve hatta bir örnekteki bilginin konuya olan ilgi derecesini de belirleyebilmelidir.

Diğer önemli bir özellik de öğrenme aşaması sonunda geliştirilen modelin öğrenilen konunun uzmanları tarafından anlaşılıp kontrol edilebilir olmasıdır. Bu özelliğin iki ayrı önemi söz konusudur. Birincisi uzmanlar tarafından denetlenebilen bir model daha güvenli bir şekilde uygulamaya konulabilir. Diğer bir yararı da uzmanların daha önceden dikkatini çekmemiş olan bazı gerçeklerin de ortaya

çıkabilmesidir. Bu özelliğe sahip öğrenme programları çeşitli konularda ilginç keşiflere de yol açmıştır.

Daha önce geliştirilmiş olan makina öğrenmesi algoritmaları yukarıda sıralanan özelliklerden ancak bir kısmına sahiptir. Örneğin bazı algoritmalar çok hızlı bir şekilde bir model öğrenmelerine rağmen, bu modeli kullanarak karar vermek çok zaman alabilmektedir (Aha, Kibler, Albert, 1991). Diğer taraftan bazı algoritmalar hızlı öğrenip hızlı karar verebilmelerine karşın verdikleri kararlar çok doğru olmamakta ya da üretilen model insanlar tarafından kolaylıkla anlaşılamamaktadır (Holte, 1993; Quinlan, 1986).

Yukarıda sıralanan özelliklerin gerçekleştirilebilmesi, temelde, öğrenilen bilginin ne şekilde gösterileceği, diğer bir deyişle kullanılan bilgi gösterim tekniği ile doğrudan ilgilidir. Bütün bu özellikleri sağlayacak şekilde bilgi gösterim tekniklerinin geliştirilmesi yönünde Bilkent üniversitesi Bilgisayar ve Enformatik Mühendisliği Bölümünde çeşitli araştırmalar sürdürülmektedir (Akkuş, Güvenir, 1996; Demiröz, Güvenir ve İlter, 1997; Demiröz, Güvenir, 1997; Demiröz, Güvenir, 1996; Güvenir, Acar, Demiröz, Çekin, 1997; Güvenir, Şirin, 1996). Bu çalışmalarda, hem hızlı öğrenilen hem de hızlı ve yüksek doğrulukta kararlar verebilen, örneklerdeki hatalardan ve eksik bilgilerden az etkilenen, ve bunun yanında örneklerdeki bilgilerin öğrenilen konuya olan ilgilerinin de öğrenilebildiği bilgi gösterim teknikleri geliştirilmiştir.

Öznitelik izdüşümü tabanlı gösterim adı verilen bu gösterimde öğrenme örnekleri sabit sayıda öznitelik ile tanımlanır. Bir özniteliğin değeri bir hastanın yaşında olduğu gibi bir tamsayı olabileceği gibi kanındaki şeker oranında olduğu gibi bir reel sayı da olabilir. Hatta, bir bölgenin rengini gösteren bir öznitelik beyaz, kırmızı, mor, sarı gibi nominal değerler de alabilir. Bu gösterim tekniğinde önce tüm öğrenme örneklerinin ayrı ayrı her bir öznitelik üzerindeki izdüşümleri alınır. Daha sonra birbirine benzer özelliklerdeki izdüşüm noktaları ayrı ayrı gruplandırılır. Yeni bir durum karşısında karar verirken önce yeni durumun her bir öznitelik açısından, ayrı ayrı izdüşümlerinin hangi aralık gruplarına düştüğü belirlenir. Her bir öznitelik için sadece bulunan aralık grubuna bakarak bir karar verilir. Daha sonra, özniteliklerin kendi kararları arasında bir oylama yapılarak son karar verilir. Değeri bilinmeyen öznitelikler oylamaya katılmazlar.

Yapılan deneyler göstermiştir ki, bu bilgi gösterim tekniği ile çok hızlı bir şekilde ve yüksek doğruluk derecelerinde kararlar alınabilmektedir. Ayrıca model oldukça hızlı bir şekilde öğrenilmektedir. Değeri belli olmayan öznitelikler oylamaya katılmadığı için karar verme aşamasında hiç bir engel teşkil etmemektedirler. Hatalı bir örneğin yol açabileceği yanlış bir karar, tüm öznitelik kararları arasında oylama yapıldığı için, doğru karar veren çoğunluğun kararları arasında etkisini kaybetmektedir. öğrenilen bilgi ayrı ayrı öznitelik değerlerine karşı düşen değerler cinsinden tutulduğu için de uzmanlar tarafından kolaylıkla anlaşılıp kontrol edilebilmektedir.

Geliştirilen bu bilgi gösterim tekniği ve ilgili öğrenme ve karar verme algoritmaları tıbbın iki ayrı dalında uygulanmıştır.

Birinci uygulama alanı dermatoloji olmuştur (Demiröz, Güvenir ve İlter, 1997). Dermatoloji kliniklerinde oldukça sık karşılaşılan ve Erythemato-squamous hastalıkları adı verilen altı hastalık mevcuttur. Bu hastalıkların doktor tarafından kolaylıkla gözle görülen klinik özellikleri birbirlerine çok benzemektedir. Bu nedenle de Erythemato-squamous hastalıklarının birbirinden ayrılması dermatolojide önemli problemlerden birisidir. Eritem ve pullanmanın derecesi, lezyonların sınırlarının belirgin olup olmaması, kaşıntının ve kobner olayının varlığı, papüllerin şekli, ağız mukozasının, dirseklerin, diz ve kafa derisinin etkilenmesi ve hatta ailede bu tür bir hastalığın görülmüş olup olmaması ve hastanın yaşı gibi klinik bulgular (öznitelikler) genellikle tanı koymak için yeterli olmamaktadır.

Bu durumda uzmanlar hastalıklı bölgelerden deri örnekleri alıp biyopsi yapma gereği duymaktadırlar. Bu biyopsi sırasında çeşitli histopatolojik özniteliğin varlığı ve varsa derecesi belirlenir.

Gazi Üniversitesi Tıp Fakültesi Dermatoloji Bölümününde gerçek hastalardan alınan bilgiler ile oluşturulan 366 örnek durum öğrenme algoritmasına verilmiş ve yukarıda belirtilen altı hastalığı birbirinden ayırmada kullanılacak bir model öğrenilmiştir. Bu modeli kullanan karar verme algoritması yeni bir hastaya ait bilgiler verildiğinde %96 oranında doğru tanı koymaktadır.

Geliştirilen bir genetik algoritma sayesinde bu 34 özniteliğin göreceli ağırlıkları da öğrenilmiştir (Demiröz, Güvenir, 1996). Bu öznitelik ağırlıklarının da karar vermede kullanılmasıyla doğruluk oranı %98'i geçmiştir.

Diğer bir uygulama alanı da EKG sinyallerinin incelenerek kalpteki ritim bozukluklarının teşhis edilmesi olmuştur (Güvenir, Acar, Demiröz, Çekin, 1997). Burada hastanın vücuduna takılan 9 elektrod yardımıyla alınan elektrik sinyallerinden üretilen 12 standart EKG işareti önce bilgisayar tarafından işlenerek 275 özniteliğin değeri belirlenmektedir. Belirlenen öznitelikler nabız, her 12 kanal için P ve Q tepeleri arasındaki süre, P, Q, R, S ve T dalgalarının süreleri, ve genlikleri gibi ölçülen değerler yanında hastanın yaşı, boyu, ağırlığı ve cinsiyetinden oluşmaktadır.

Başkent üniversitesi Tıp Fakültesi uzmanlarından Doç. Dr. Haldun Müderrisoğlu ve Dr. Ayhan Çekin tarafından ve gerçek hastalardan alınan bilgiler ile oluşturulan 452 örnek kayıt öğrenme algoritmasına verilmiş ve ritim bozukluğu olup olmadığını belirlemede ve varsa 15 farklı ritim bozukluğunu birbirinden ayırmada kullanılacak bir model öğrenilmiştir. Bu modeli kullanan karar verme algoritması yeni bir hastaya ait bilgiler verildiğinde %62 oranında doğru tanı koymaktadır.

Benzer şekilde geliştirilen bir genetik algoritma sayesinde bu 279 özniteliğin göreceli ağırlıkları da öğrenilmiştir. Bu öznitelik ağırlıklarının da karar vermede kullanılmasıyla doğruluk oranı %67'e yaklaşmıştır.

Geliştirilen sınıflandırma ve ağırlık öğrenme algoritmaları yukarıda belirtilen verikümeleri dışında, makina öğrenmesi araştırmalarının karşılaştırılmasında bir standart oluşturmak amacıyla derlenmiş bulunan UCI-Repository`deki verikümeleri üzerinde de sınanmıştır. Geliştirilen bu algoritmalar standart verikümeleri üzerinde de çok başarılı sonuçlar elde etmişlerdir. Bu araştırmalar göstermiştir ki geliştirilen bilgi gösterim teknikleri ve ilgili makina öğrenmesi teknikleri birçok alanda başarı ile uygulanabilecektir.

Proje, öneride tanımlanan amaç doğrultusunda gerçekleştirilmiş, öznitelik bölümlemesi ve çakışan öznitelik aralıkları şeklinde çoklu kavram tanımlamalarının toplu olarak öğrenimine ve öznitelik ağırlıklarının genetik algoritmalar kullanılarak belirlenmesine yönelik teknikler geliştirilmiştir.

Proje önerisinde tanımlandığı şekilde, proje amaçlanan algoritma ve tekniklerin geliştirilip gerçekleştirilmesini, biri dermatoloji bir diğeri de EKG işaretlerinden aritmi sınıfının belirlenmesi olmak üzere iki verikümesinin oluşturulmasını ve bu algorirmalar üzerinde sınanmasını kapsamaktadır.

Projenin gerçekleştirilmesinde kullanılan yöntem de öneride belirtildiği şekildedir. Geliştirilen algoritmalar proje bütçesinden alınan bir SUN Sparc 20/61 bilgisayar üzerinde ve C dilinde gerçekleştirilmiş ve derlenen veri kümeleri üzerinde sınanmıştır. Proje bütçesi 1,154,025,000 TL olup, bu meblağın 1,154,000,000 TL tutarındaki kısmı ile, proje önerisinde belirtildiği şekilde bir SUN Sparc 20/61 bilgisayar satın alınmıştır.

Proje önerisinde geliştirilen algoritmaların bu veri kümeleri üzerinde en az yüzde seksen doğrulukta sınıflandırma yapmaları hedeflenmişti. Geliştirilen algoritmalar dermatoloji veri kümesi üzerinde %98 doğrulukta sınıflandırma yapabilmektedir. Aritmi veri kümesi üzerinde ise %67 doğrulukta sınıflandırma yapılabilmektedir. Bunun nedeni ise EKG işaretleri alınan hastalarının çoğunun normal (aritmi yok) olması, aritmili hastaların ise genellikle birkaç aritmi sınıfında olmaları, hatta bazı aritmi türlerine ise proje süresince hiç karşılaşılmamış olamasıdır. Buna rağmen %67 doğrulukta sınıflandırma yapılabilmesi yüksek bir başarı olarak değerlendirilmelidir.

Raporun ikinci bölümünde konu ile ilgili daha önceki çalışmalar özetlenecektir. Proje çerçevesinde derlenen dermataoloji ve aritmi verikümelerinin detayları Üçüncü Bölümde verilmektedir. Raporun Gelişme Bölümünde geliştirilen bilgi gösterim teknikleri ve bunları kullanan öğrenme, sınıflandırma ve agırlık öğrenme algoritmaları detaylı olarak açıklanacak ve çeşitli verki kümeleri üzerinde elde edilen sonuçlar diğer benzer algoritmalarla karşılaştırmalı olarak verilecektir. Ayrıca oluşturulan iki tıbbi verikümesi detaylı olarak açıklanacaktır. Raporun Sonuç Bölümü ise sonuçlar kısaca sıralanarak irdelenecektir. Ayrıca ileriye dönük araştırma önerileri sunulacaktır.

2 DAHA ÖNCEKİ ÇALIŞMALAR

Proje çerçevesinde gerçekleştirilen çalışma, verilen çok sayıda sınıflandırılmış örnekten yola çıkarak ve çeşitli genelleme yöntemleri kullanarak yeni örneklerin (durumların) doğru birşekilde sınıflandırılmasına yönelik tekniklerin geliştirilmesidir.

Örneklerden sınıflandırma tekniklerinin geliştirilmesi konusundaki ilk kapsamlı çalışma Duda ve Hart (1973) tarafından yapılmıştır. Bu çalışmada Bayesian sınıflandırıcı ve en yakın komşu (nearest neighbor) sınıflandırıcısı tanımlandırılmıştır. Daha sonra Medin ve Schaffer (1978) insanlarda sınıflandırmanın öğrenilmesi konusunda yaptıkları çalışmada, insanların kavramları karekteristik örnekler (examplars) şeklinde öğrendiklerini iddia etmişlerdir. Aha ve diğerleri daha sonra bu fikri makina öğrenmesine uygulayarak IB1, IB2 ve IB3 adlı artımsal (incremental) vakaya dayalı (Insatance-Based) öğrenme teknikleri geliştirmiştir (Aha, 1989; Aha, Kibler, Albert, 1991). Wettschereck ve Aha (1995) daha örnek tabanlı ve en yakın komşu sistemlerinde özniteliklerin önem derecelerinin belirlenmesi üzerinde çalıştılar. Salzberg karekteristik örnekleri genelleyerek iç içe geçmiş hiper-dikdörtgenler şeklinde gösterim metodunu önermiştir (Salzberg, 1991; Salzberg, 1990). Daha sonra Güvenir ve Şirin örneklerin öznitelikler üzerindeki izdüşümleri cinsinden gösterilmesini ve sınıflamanın da özniteliklere göre yapılacak sınıflamalar arasında bir oylama ile yapılmasını önerdi (Güvenir, Şirin, 1993; Şirin, Güvenir, 1993). Bu metoda CFP (Classification by Feature Partitioning) adı verildi.

De Jong (1988) genetik algoritmaların mekanik öğrenmede kullanılnasını inceledi. Goldberg (1989) parametre optimizasyonunda ve makina öğrenmesinde genetik algoritmaların kullanılması üzerine çeşitli çalışmalar yaptı. Kelly ve Davis (1991) sınıflandırmada öznitelik ağırlıklarının belirlenmesi için genetik algoritmaların kullanılmasını önerdiler. Güvenir ve Şirin (1993) CFP'de öznitelik ağırlıklarının belirlenmesinde genetik algoritmalardan yararlandılar. Holte (1993) sınıflamanın sadece en ayırıcı özniteliğe göre yapıldığında bile (ki bu CFP'nin bir özel haline karşı düşmektedir) başarısının fazla düşmediğini gösterdi. Son olarak Ünsal-Koç (1995) çakışan öznitelik aralıkları şeklinde tanımlama bilgisinin artımsal olarak öğrenimi için COFI (Classification with Overlapping Feature Intervals) adı verilen yöntemi önerdi.

Makina öğrenmesinin tıp alanında arasında da literatürde çeşitli yayınlara rastlanmaktadır. Spackman (1988) biomedikal alanlarda kategorik karar kriterlerininin mekanik olarak öğrenilmesi konusunda ilk çalışmalardan birini (CRLS: Criteria Learning System) gerçekleştirmiştir. CRLS adlı sistem bir tanı koymak için *n* öznitelikten *k* tanesinin gözlenmiş olmasını esas alır. Kononenko (1996) tümevarım temelli ve Bayesiyan öğrenme tekniklerinin tıbbi tanı koyma uygulamalarının güzel bir özetini vermektedir.

3 DERLENEN VERİKÜMELERİ

Bu proje çerçevesinde geliştirilen öğrenme ve sınıflandırma algoritmalarının sınanması ve uygulanmasına yönelik olarak, tıbbın iki farklı alanında birer verikümesi oluşturulmuştur. Bu verikümelerinden birincisi, dermatoloji kliniklerinde oldukça sık karşılaşılan ve erythemato-squamous hastalıklarının tanınmasına yönelikdir. Raporun bundan sonraki bölümlerinde bu verikümesi *dermatoloji aritmi* (ya da *dermatoloji*) *verikümesi* olarak adlandırılacaktır. Derlenen ikinci verikümesi ise EKG sinyallerinin incelenerek kalpteki ritim bozukluklarının teşhis edilmesine yöneliktir. Bu verikümesi *ise aritmi* (ya da *arrhythmia*) *verikümesi* olarak anılacaktır. Bu verikümelerinin detayları aşağıdaki altbölümlerde verilmiştir.

Dermatoloji ve Aritmi verikümeleri UCI-Repository'deki formatta hazırlanmıştır. Verikümelerinin özelliklerini içeren `dermatology.names` ve `arrhythmia.names` (Bkz. Ek-2 ve Ek-3) ve verileri içeren `dermatology.data` ve `arrhythmia.data` adlı dosyalar isteyen araştırmacılara gönderilecektir.

Dermatoloji Verikümesi

Dermatoloji kliniklerinde oldukça sık karşılaşılan ve Erythemato-squamous hastalıkları adı verilen altı hastalık mevcuttur. Bu gruptaki hastalıklara *psoriasis*, *seberoic dermatitis*, *lichen planus*, *pityriasis rosea*, *cronic dermatisis* ve *pityriosis rubra pilaris* adları verilmektedir. Dermatoloji verikümesine kullanılan sınıflar ve örnek sayıları Tablo 1`de gösterilmiştir. Bu hastalıkların doktor tarafından kolaylıkla gözle görülen klinik özellikleri birbirlerine çok benzemektedir. Bu nedenle de Erythemato-squamous hastalıklarının birbirinden ayrılması dermatolojide önemli problemlerden birisidir. Eritem ve pullanmanın derecesi, lezyonların sınırlarının belirgin olup olmaması, kaşıntının ve kobner olayının varlığı, papüllerin şekli, ağız mukozasının, dirseklerin, diz ve kafa derisinin etkilenmesi ve hatta ailede bu tür bir hastalığın görülmüş olup olmaması ve hastanın yaşı gibi klinik bulgular (öznitelikler) genellikle tanı koymak için yeterli olmamaktadır. Bu durumda uzmanlar hastalıklı bölgelerden deri örnekleri alıp biyopsi yapma gereği duymaktadırlar. Bu biyopsi sırasında çeşitli histopatolojik özniteliğin varlığı ve varsa derecesi belirlenir.

Tablo 1. Dermatoloji verikümesindeki sınıflar ve örnek sayıları.

Sınıf	Örnek sayısı
C_1: psoriasis	112
C_2: seboreic dermatitis	61
C_3: lichen planus	72
C_4: pityriasis rosea	49
C_5: cronic dermatitis	52
C_6: pityriasis rubra pilaris	20

Gazi üniversitesi Tıp Fakültesi öğretim üyelerinden Doç. Dr. Nilsel İlter tarafından ve gerçek hastalardan alınan bilgiler ile oluşturulan verikümesinde 366 adet örnek durum (veri) bulunmaktadır. Her hasta ile ilgili olarak 34 öznitelik değeri alınmaktadır. Bunlardan 12 tanesi doktorun çıplak gözle belirleyebildiği klinik öznitelikler, gerikalan 22 tanesi ise laboratuvarda mikroskop altında belirlenen histopatolojik özniteliklerdir. Dermatoloji verikümesine kullanılan öznitelikler Tablo 2`de gösterilmiştir. Dermatoloji verikümesinin detaylarını içeren dermatology.names adlı dosya Ek-2 de verilmiştir.

Tablo 2. Dermatoloji verikümesindeki öznitelikler.

Öznitelikler	
Klinik Öznitelikler	Histopatolojik Öznitelikler
f_1: erythema	f_{12}: melanin incontinence
f_2: scaling	f_{13}: eosinophils in the infiltrate
f_3: definite borders	f_{14}: PNL infiltrate
f_4: itching	f_{15}: fibrosis of the papillary dermis
f_5: koebner phenomenon	f_{16}: exocytosis
f_6: plygonal papules	f_{17}: acanthosis
f_7: follicular papules	f_{18}: hyperkeratosis
f_8: oral mucosal involvement	f_{19}: parakeratosis
f_9: knee and elbow involvement	f_{20}: clubbing of the rete ridges
f_{10}: scalp involvement	f_{21}: elongation of the rete ridges
f_{11}: family history	f_{22}: thinning of the suprapapillary epidermis
f_{34}: age	f_{23}: spongiform pustule
	f_{24}: munro microabcess
	f_{25}: focal hypergranulosis
	f_{26}: disappearence of granular layer
	f_{27}: vacuolization and damage of basal layer
	f_{28}: spongiosis
	f_{29}: saw-tooth appearance of retes
	f_{30}: follicular horn plug
	f_{31}: perifollicular parakeratosis
	f_{32}: inflammatory mononuclear inflitrate
	f_{33}: band-like infiltrate

Bu özniteliklerden f_{11} (family history) ailede bu hastalıklardan biri daha önce görülmüş ise 1, aksi halde 0 değerini almaktadır. Yaş (f_{34}) hastanın yaşını gösteren

bir tam sayıdır. Diğer öznitelikler ise 0, 1, 2 veya 3 değerini almaktadır. Eğer sözkonusu öznitelik hastada hiç mevcut değilse 0, çok belirgin bir şekilde gözlendi ise 3 değerini almaktadır. 18 yaşındaki bir cronic dermatitis hastasına ait bilgi aşağıdaki şekilde tutulmaktadır.

```
2,1,0,2,0,0,0,0,0,0,0,0,0,3,1,3,0,0,0,2,0,0,0,0,0,0,0,0,0,0,2,0,18,5
```

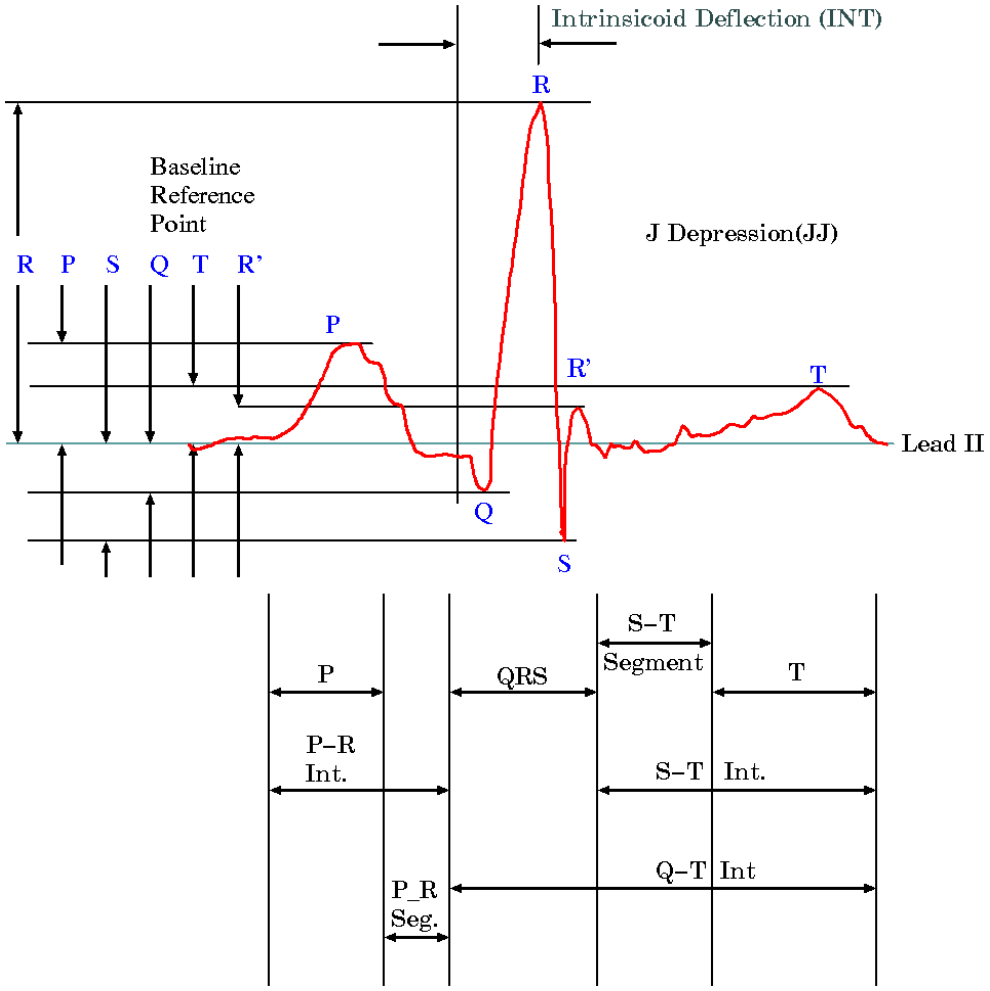
Bu örnekte lezyon sınırları hiç belirgin değildir (f_3 = 0), buna rağmen papillary dermisde çok belirgin bir şekilde fibrosis görülmüştür (f_{15} = 3).

Aritmi Verikümesi

Aritmi verikümesi EKG sinyallerinin incelenerek kalpteki ritim bozukluklarının teşhis edilmesi amacıyla oluşturulmuştur. Burada hastanın vücuduna takılan 9 elektrod yardımıyla alınan elektrik sinyallerinden üretilen 12 standart EKG işareti önce bilgisayar tarafından işlenerek 279 özniteliğin değeri belirlenmektedir. Bu 12 EKG işaretinden birinin tipik bir şekli Şekil 1'de gösterilmiştir. Belirlenen öznitelikler nabız, her 12 kanal için P ve Q tepeleri arasındaki süre, P, Q, R, S ve T dalgalarının süreleri, ve genlikleri gibi ölçülen değerler yanında hastanın yaşı, boyu, ağırlığı ve cinsiyetinden oluşmaktadır. 279 özniteliğin 206'sı lineer, geri kalan 73'ü ise nominaldir. Aritmi verikümesinin detaylarını içeren arrhythmia.names adlı dosya Ek-3 de verilmiştir.

Başkent üniversitesi Tıp Fakültesi uzmanlarından Doç. Dr. Haldun Müderrisoğlu ve Dr. Ayhan Çekin tarafından ve gerçek hastalardan alınan bilgiler Bilkent Üniversitesi Elektrik ve Elektronik Mühendisliği Bölümü araştırma görevlilerinden Burak Acar tarafından işlenerek her hasta için öznitelik değerleri belirlenmiştir. Bu verikümesinde 452 örnek kayıt bulunmaktadır. Gerçekte tıpta 20 farklı aritmi çeşiti tanımlanmaktadır. Bunlardan bir tanesi her hangi bir aritminin olmadığı normal (C_1) durumdur. Beş tip aritmi çeşidi benzer özellikleri nedeniyle tek bir grup altında (C_{16}) toplanmaktadır. Aritmi verikümesindeki sınıflar ve örnek sayıları Tablo 3`de gösterilmiştir.

Tablo 3`den de görüleceği üzere C_{11}, C_{12} ve C_{13} örneklemenin yapıldığı grupda hiç gözlenmemiştir. Bunun yanında örneklerin yaklaşık yarısı normal sınıfa (C_1) aittir. Bu özelliğinden dolayı aritmi verikümesi öğrenilmesi zor bir verikümesidir.



Şekil 1. Tipik bir EKG işareti.

Tablo 3. Aritmi verikümesindeki sınıflar ve örnek sayıları.

Sınıf	Açıklama	Örnek Sayısı
C_1	Normal	245
C_2	Ischemic changes (KAH) (Coronary Artery Disease)	44
C_3	Old Anterior Myocardial Infarction	15
C_4	Old Inferior Myocardial Infarction	15
C_5	Sinus tachycardy	13
C_6	Sinus bradycardy	25
C_7	Ventricular Premature Contraction (PVC)	3
C_8	Supraventricular Premature Contraction	2
C_9	Left bundle branch block	9
C_{10}	Right bundle branch block	50
C_{11}	1. degree AtrioVentricular block	0
C_{12}	2. degree AV block	0
C_{13}	3. degree AV block	0
C_{14}	Left ventricule hypertrophy	4
C_{15}	Atrial Fibrillation or Flutter	5
C_{16}	Diğer aritmiler	22

4 ÖZNİTELİK BÖLÜMLEMESİ ŞEKLİNDE ÇOKLU KAVRAM TANIMLARININ ÖĞRENİLMESİ

Proje kapsamında geliştirilecek olan öznitelik bölümlemesi (feature partitioning) ve çakışan öznitelilk aralıkları (overlapping feature intervals) şeklinde çoklu kavram tanıma algoritmalarına bir karşılaştırma olanağı sağlamak amacıyla, literatürde yaygın olarak karşılaşılan *k* en yakın komşu sınıflandırması (*k*-NN, *k* Nearest Neighbor) algoritması öznitelik izdüşümleri üzerinde sınıflandırma yapacak şekilde, *k*-NNFP (*k* Nearest Neighbor Classification on Feature Projections) adıyla yeniden tanımlanmıştır. Bu algoritma UCI-Repository'deki (University of California at Irvine) doğal veri kümeleri, ve ayrıca çeşitli sayılarda ilgisiz (irrelevant) öznitelikler içeren yapay veri kümeleri üzerinde sınanmış, sonuçlar "K Nearest Neighbor Classification on Feature Projections" adlı bir bidiri olarak 13. Uluslararası Machine Learning Konferansında sunulmuş ve bildiriler kitabında yayınlanmıştır. Bildirinin bir kopyası Ek-4'de verilmiştir. Bu bildiride ayrıca *k*-NNFP algoritması yaygın olarak bilinen *k*-NN algoritmasıyla karşılaştırılmıştır. Bu karşılaştırmalarda, öznitelik izdüşümleri üzerinde yapılacak olan sınıflandırmaların özellikle veri kümesi içinde ilgisiz özniteliklerin olduğu durumlarda *k*-NN algoritmasından daha başarılı olduğu gözlemlenmiştir.

Projenin ilk aşamasında, proje önrisinde olduğu gibi, öznitelik bölümlemesi şeklinde çoklu kavram tanımları (BCFP, Batch Classification on Feature Projections) algoritmaları geliştirilmiştir. Bu algoritmaların temel yapısı Şekil 2'de verilmiştir. Bu algoritmalar tarafından öğrenilen kavram tanımları, her öznitelik için öğrenilen bölümlemeler şeklinde verilir. Dolayısıyla, bir kavram tanımı, her öznitelik için, sözkonusu kavrama ait olan değer aralıklarından (bölüntülerden) oluşur. Bu algoritmalarda sınıflandırma, öznitelikler arasındaki bir ağırlklı oylama ile belirlenir.

Bu bilgi gösterimi yönteminde verilen bir özniteliğin belirli bir değeri için, normal olarak, sadece tek bir sınıfa oy verilir. Ancak eğer öğrenme kümesi içinde bu değer için farklı sınıflara ait örnekler mevcutsa bu örneklerin nasıl saklanacağı ve sınıflamada nasıl kullanılacağı sorusuna farklı yanıtlar verilebilir. Bu farklı yanıtlar aşağıdaki farklı BCFP algoritmalarını ortaya çıkarmıştır.

```
öğren(ÖğrenmeKümesi)
başlangıç
   ÖğrenmeKümesi'ni sırala            /* bütün öznitelik izdüşümleri üzerinde */
   öznitelikBölüntüleriniOluştur (ÖğrenmeKümesi)
bitiş

öznitelikBölüntüleriniOluştur(ÖğrenmeKümesi)
başlangıç
   her öznitelik f için
      ÖğrenmeKümesi içindeki her öğrenme örneği i için
         noktaBölüntüleriOluştur (f, i)
      eğer öznitelik f sayısal değerler alıyorsa
         bölüntüleriGenelle (f )
      /* Nominal değerli öznitelikler için genelleme yapılamaz */
bitiş

bölüntüleriGenelle(f)
başlangıç
   her ardışık bölüntü çifti i ve j için
      eğer i ve j bölüntüleri aynı sınıftan olan nokta bölüntüler ise
         i ve j bölüntülerini tek bir bölüntü halinde birleştir  /* genelle */
bitiş

sınıfla(örnek)
başlangıç
   her sınıfın oyunu 0'a eşitle
   her öznitelik f için
      /* f özniteliği üzerinde, örneğin f değerine karşı gelen bölüntüyü bul */
      bölüntü = bölüntüAra(f, örnek_f)
      eğer bölüntü'nün sınıfı belirli ise
         bölüntü'nün sınıfının oyunu 1 arttır
   tahmin = belirsiz
   her sınıf c için
      eğer c'nin oyu tahmin'in oyundan büyük ise
         tahmin = c
   tahmin'i döndür
bitiş
```

Şekil 2. BCFP Algoritmalarının genel yapısı.

BCFP algoritmaları C dilinde yazılıp, gcc (GNU C Compiler) derleyicisiyle derlenip, proje kapsamında satın alınan iş istasyonu üzerinde çalıştırılmıştır.

BCFP.1:

Bu projede kullanılan verikümelerinin hepsinde her bir örnek bir vektör ile gösterilir. Vektörün ilk n değeri n özniteliğin değerlerini, sonuncu değer ise örneğin sınıfını gösterir. BCFP1 algoritmasında bir bölüntü ise dört değerli bir vektör ile tanımlanır. Bunların ilk ikisi bölüntünün alt ve üst sınırlarını, üçüncü değer bölüntünün temsil ettiğiörnek sayısını, sonuncu değer ise bölüntünün sınıfını (oy verdiği sınıfı) belirtir.

Örnek olarak aşağıdaki üç öznitelikli ve üç sınıflı öğrenme verikümesinin verildiğini varsayalım:

i_1=<1,10,7,C_1> i_{10}=<10,17,16,C_2>

i_2=<3,12,7,C_1> i_{11}=<10,17,18,C_2>

i_3=<4,10,10,C_1> i_{12}=<4,17,4,C_3>

i_4=<4,12,10,C_1> i_{13}=<4,17,4,C_3>

i_5=<4,15,10,C_1> i_{14}=<4,17,1,C_3>

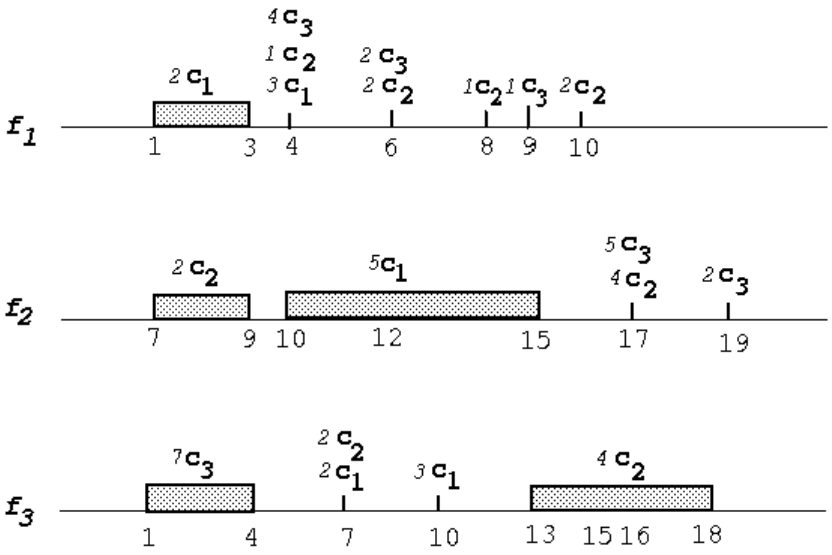
i_6=<4,7,7,C_2> i_{15}=<4,17,4,C_3>

i_7=<6,17,7,C_2> i_{16}=<6,17,4,C_3>

i_8=<6,9,13,C_2> i_{17}=<6,19,1,C_3>

i_9=<8,17,15,C_2> i_{18}=<6,19,4,C_3>

Bu örnek verikümesi için öğrenilen öznitelik bölüntileri Şekil 3'de gösterilmiştir.



Şekil 3. BCFP.1 algoritmasıyla oluşturulan örnek öznitelik bölüntüleri.

```
sınıfla(örnek)
başlangıç
   her sınıfın oyunu 0'a eşitle
   her öznitelik f için
   /* f özniteliği üzerinde, örneğin f değerine karşı gelen bölüntüyü bul */
   bölüntü = bölüntüAra(f, örnekf )
      eğer bölüntü'nün sınıfı belirli ise
         bölüntü'nün sınıfının oyunu f özniteliğinin ağırlığı kadar artır
   tahmin = belirsiz
   her sınıf c için
      eğer c'nin oyu tahmin'in oyundan büyük ise
         tahmin = c
   tahmin'i döndür
bitiş
```

Şekil 4. BCFP.1 algoritmasının sınıflandırma yordamı.

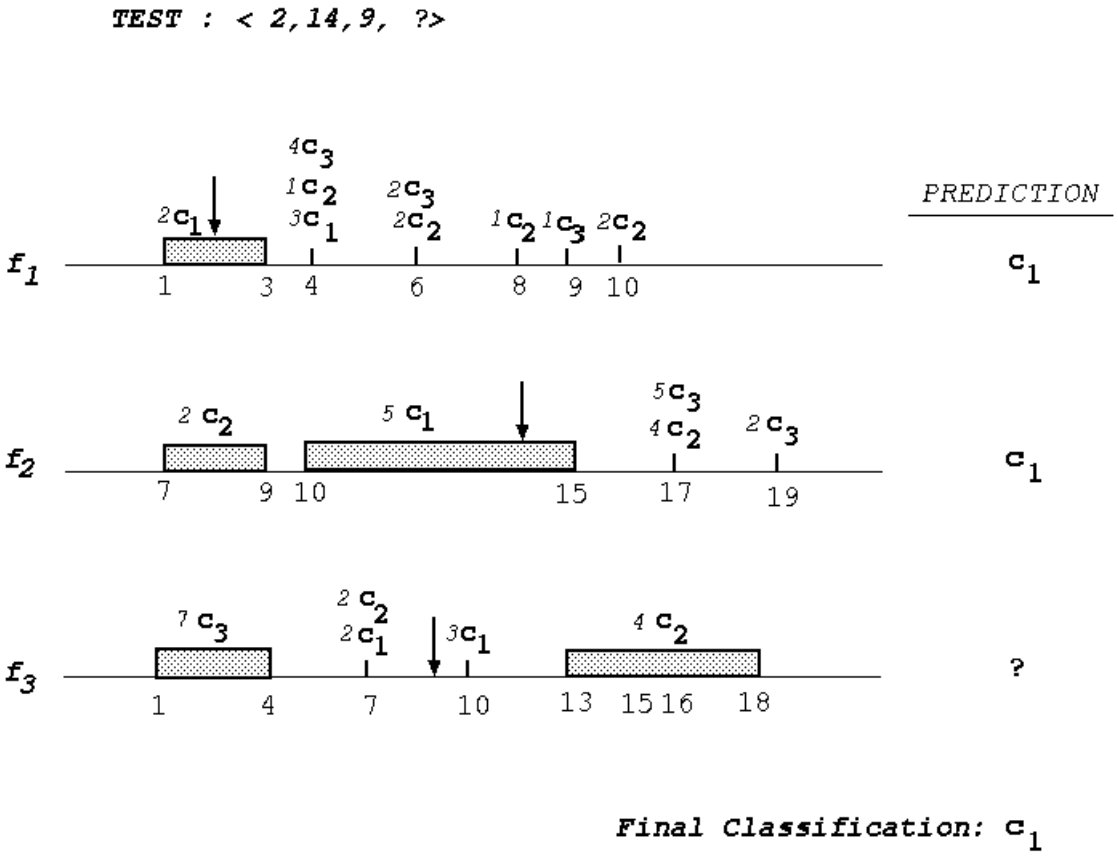
Bu durumda oluşturulan bölüntüler şunlardır:

<1,3,2,C_1> <4,4,3,C_1> <4,4,1,C_2>
<4,4,4,C_3> <6,6,2,C_2> <6,6,3,C_3>
<8,8,1,C_3> <9,9,1,C_3> <10,10,2,C_2>

BCFP.1 algoritmasında, bir özniteliğin oyunun belirlenmesi sırasında eğer birden fazla değer varsa, ki bu durum nominal değerli öznitelikler için sıkça karşılaşılan bir durumdur, bu değerde ençok örneği bulunan sınıf kullanılmaktadır. Bu durum öğretme kümesinde sınıfların dağılımının eşit olmadığı durumlarda az örneği bulunan sınıfların doğru tahmin edilme şansını azaltmaktadır. BCFP.1 algoritmasının sınıflandırma yordamı Şekil 4'de verilmiştir.

Bu sınıflandırma algoritmasına göre, Şekil 3'deki öznitelik bölüntüleri kullanılarak <2,14,9,?> yeni durumunun sınıflandırılması Şekil 5'de gösterilmiştir. Şekilden de görüleceği üzere f_1 ve f_2 öznitelikleri yeni durumun sınıfını C_1 olarak tahmin etmektedir. f_3 özniteliği ise bu durumda hiç bir tahminde bulunmamaktadır. Böylece, bu yeni durum C_1 olarak sınıflandırılmaktadır.

Öznitelik ağırlıklarının eşit değerde olduğu varsayılarak çalıştırılan BCFP.1 programının UCI-Repository'deki bazı doğal verikümeleri ve bu proje çerçevesinde derlenen dermatoloji ve aritmi verikümeleri için elde edilen 5-katlı çapraz-sağlama (5-fold cross-validation) sınıflandırma doğruluğu Tablo 4'de verilmiştir.



Şekil 5. BCF1 algoritmasıyla örnek bir sınıflandırma.

BCFP.2:

BCFP.1 algoritmasında, öğrenme kümesindeki dağılım farklılıklarından kaynaklanan problemleri önleme amacıyla, öznitelik oyunun belirlenmesi sırasında her sınıfın oy şansı o sınıftan olan tüm örneklerin sayısına oranlanarak normalize edilmesiyle BCFP.2 algoritması elde edilmiştir. BCFP.1 de kullanılan bölüntüAra yordamı Şekil 6'de, BCFP.2 de kullanılan bölüntüAra yordamı ise Şekil 7'de verilmiştir.

```
bölüntüAra (f, v)
başlangıç
   f özniteliği üzerinde v değerini kapsayan tüm bölüntüleri bul
   bu bölüntüler arasından örnek sayısı ençok olan bölüntüyü döndür
bitiş
```

Şekil 6. BCFP.1 algoritmasında kullanılan bölüntüAra yordamı.

bölüntüAra (f, v)
başlangıç
  f özniteliği üzerinde v değerini kapsayan tüm bölüntüleri bul
  bu bölüntüler arasından
    $\left(\frac{\text{bölüntüdeki örnek sayısı}}{\text{bu sınıftan toplam örnek sayısı}} \right)$ oranı ençok olan bölüntüyü döndür
bitiş

Şekil 7. BCFP.2 algoritmasında kullanılan bölüntüAra yordamı.

Yine öznitelik ağırlıklarının eşit olduğu varsayılarak çalıştırılan BCFP.2 programının aynı verikümeleri için elde edilen 5-katlı çapraz-sağlama sınıflandırma doğruluğu Tablo 4'de verilmiştir. Tablodan da görüleceği üzere verikümesinde sınıf dağılımlarının farklılık gösterdiği cleveland, dermatology, horse, hungarian musk ve wine verikümelerinde BCFP.2 algoritması BCFP.1 algoritmasına oranla daha başarılı olmuştur. Diğer veri kümelerinde ise belirgin bir fark görülmemektedir.

BCFP.3:

BCFP.1 ve BCFP.2 algoritmalarının başarısı dışarıdan verilen öznitelik ağırlıklarının doğruluğuna bağımlıdır. Eğer öznitelik ağırlıkları doğru olarak verilmiş ise bu algoritmalar da yüksek bir doğruluk elde edebileceklerdir. Aksi halde doğruluk da düşük olacaktır. Dolayısıyla bu algoritmaların başarısı için doğru öznitelik ağırlıklarının da öğrenilmesi gerekecektir. Diğer yandan BCFP.1 ve BCFP.2 algoritmalarında öznitelik ağırlıkları ait oldukları özniteliğin tüm değerleri için aynı kalmaktadır. Ancak dermatoloji ve aritmi verikümelerinin oluşturulması sırasında uzmanlarla yaptığımız görüşmelerde genellikle verilen bir özniteliğin bazı değerlerinin diğer değerlerinden daha önemli ve ayırıcı olduğunu gözledik.

Hem öznitelik ağırlıklarının dışarıya bağımlılığını ortadan kaldırmak ve hem de öznitelik değerlerinin farklı ağırlıklar alabilmelerini sağlayabilmek amacıyla BCFP.3 algoritması geliştirildi. BCFP.3 algoritmasının sınıflama yordamı Şekil 8'de verilmiştir. Buradaki fark oylama sırasında kullanılan oyun ağırlığının bölüntünün ağırlığına eşit olmasıdır. BCFP.3 algoritması bölüntü ağırlıklarını da kendisi öğrenmektedir. Verien bir f özniteliği üzerindeki s bölüntüsünün ağırlığının belirlenmesinde kullanılan bölüntü_ağırlığı yordamı Şekil 9'de verilmiştir.

```
sınıfla (örnek)
başlangıç
   her sınıfın oyunu 0'a eşitle
   her öznitelik f için
   /* f özniteliği üzerinde, örneğin f değerine karşı gelen bölüntüyü bul */
   bölüntü = bölüntüAra(f, örnek_f)
      eğer bölüntü'nün sınıfı belirli ise
         bölüntü'nün sınıfının oyunu bölüntünün ağırlığı * f in ağırlığı kadar artır
   tahmin = belirsiz
   her sınıf c için
      eğer c'nin oyu tahmin'in oyundan büyük ise
         tahmin = c
   tahmin'i döndür
bitiş
```

Şekil 8. BCFP.3 algoritmasının sınıflandırma yordamı.

```
bölüntü_ağırlığı (f, s)
başlangıç
   eğer s bulunduğu değerdeki tek bölüntü ise 1 döndür
   değilse
      bu değerdeki tüm bölüntüleri tek bir b bölüntüsüne topla
      b nin sınıfı = bu değerde en sık oluşan sınıf
      b nin ağırlığı = ( bu değerde en sık oluşan iki sınıf arasındaki örnek farkı
                       / bu değerde oluşan örnek sayısı )
      b nin ağırlığını döndür
bitiş
```

Şekil 9. BCFP.3 algoritmasının bölüntü_ağırlığı yordamı.

BCFP.3 algoritmasında aynı değere sahip bölüntüler birleştirildiğinde bölüntünün ağırlığı, o değerde en sık görülen iki sınıfın örnek sayıları arasındaki farkın o değerdeki toplam örnek sayısına oranı şeklinde belirlenmektedir. Bu sayede aynı değere sahip farklı iki sınıftan eşit sayıda örnek varsa o bölüntünün ağırlığı 0 olmaktadır. Böylece bu değer için sözkonusu öznitelik oylamada çekimser kalmaktadır.

BCFP.3 algoritmasının aynı doğal verikümeleri için elde edilen 5-katlı çapraz-sağlama sınıflandırma doğruluğu Tablo 4 de verilmiştir. Buna göre bazı veri kümeleri için önceki algoritmalara oranla biraz daha iyi fakat bazı veri kümelerinde de biraz daha kötü sonuçlar elde edilmiştir, ancak önemli bir iyileşme sağlanamamıştır.

BCFP.4:

BCFP.3 algoritmasında tek bir noktada tek bir örnekten oluşmuş olan bir bölüntü, en yüksek ağırlığa (= 1) sahip olmaktadır. Oysa böyle bir bölüntünün hatalı (noisy) bir örnekten dolayı oluşmuş olma olasılığı çok yüksektir. Bu nedenle bölüntü ağırlıklarının sınıflarının toplam sayısına oranlayarak normalize eden BCFP.4 algoritması geliştirildi. BCFP.4 algoritması sınıflama sırasında öznitelik ağırlıklarını da hesaba katmaktadır.

BCFP.4 algoritmasının aynı doğal verikümeleri için elde edilen 5-katlı çapraz-sağlama sınıflandırma doğruluğu Tablo 4'de verilmiştir. BCFP.4 algoritması veri kümeleri dışında diğer algoritmalara benzer veya daha iyi sonuçlar vermiştir. Burada öznitelik ağırlıklarının eşit olduğu varsayılmıştır.

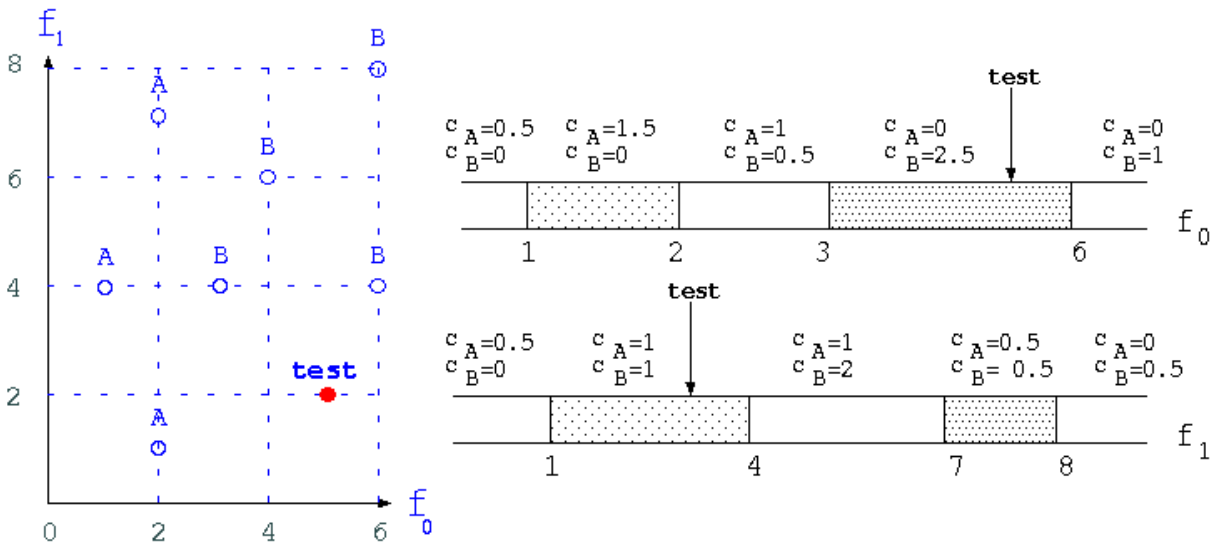
Tablo 4. BCFP.0, BCFP.1, BCFP.2, BCFP.3 ve BCFP.4 programlarının bazı doğal verikümeleri için elde edilen 5-katlı çapraz sağlama yolu ile elde edilen sınıflandırma doğrulukları (% olarak).

Veri Kümesi	BCFP.1	BCFP.2	BCFP.3	BCFP.4
arrhythmia	54.16	54.38	54.16	57.72
bcancerw	95.72	96.44	95.86	97.00
cleveland	78.22	80.52	78.86	80.50
dermatology	50.28	53.56	52.18	58.50
diabetes	65.76	64.84	68.74	69.76
glass	49.88	49.50	56.98	45.36
horse	64.12	73.64	65.24	76.36
hungarian	69.02	80.94	68.00	75.48
ionosphere	87.18	84.90	87.74	88.88
iris	86.66	88.00	90.66	90.66
liver	54.78	53.92	57.98	59.72
musk	61.96	71.04	71.02	73.34
vehicle	57.22	59.22	58.04	58.14
wine	82.54	87.62	91.60	89.92

5 ÇAKIŞAN ÖZNİTELİK ARALIKLARI ŞEKLİNDE ÇOKLU KAVRAM TANIMLARININ ÖĞRENİLMESİ

Proje önerisinde planlandığı üzere çakışan öznitelik aralıkları şeklinde çoklu kavram tanımları için toplu makina öğrenme algoritması (BCOFI, Batch Classification on Overlapping Feature Intervals) geliştirilmiştir. BCOFI algoritmasında öznitelik izdüşümleri her sınıf için ayrı ayrı olmak üzere genellenmektedir. Böylece herhangi bir öznitelik değeri için bir veya daha çok çakışan bölüntü bulunabilmektedir. Dolayısıyla, verilen bir değer için bir özniteliğin oyu tek bir sınıf için değil, o değeri içine alan bütün bölüntülerin temsil ettiği sınıflar için olabilir. Diğer bir değişle bir değer için o değeri kapsayan bütün bölüntüler içerdikleri örnek sayısı ile orantılı olarak oy kullanmaktadırlar. İki öznitelikli örnek bir verikümesi ve Çakışan Öznitelik Aralıkları şeklinde gösterimi Şekil 10'de gösterilmiştir.

BCOFI algoritmasının öğrenme ve sınıflandırma yordamları Şekil 6'da verilmiştir.



Şekil 10. İki öznitelikli bir verikümesi ve çakışan öznitelik aralıkları şeklinde gösterimi.

BCOFI Algoritmasının BCFP algorıtmalarının karşılaştırılmasında kullanılan doğal verikümeleri için elde edilen 5-katlı çapraz-sağlama sınıflandırma doğruluğu ve programın SUN-Sparc 20/61 iş istasyonundaki ortalama çalışma süreleri Tablo 5 de verilmiştir. Tabloda ayrıca, bir karşılaştırma sağlamak amacıyla, aynı ölçümler 1-NN (*k*-NN, *k*=1) algoritması için de verilmiştir. Tablodan da görüleceği üzere BCOFI algoritması 1-NN algoritmasından bir miktar daha yüksek doğrulukta sınıflandırma yapabilmekte, fakat daha da önemlisi bu sınıflandırmayı 1-NN algoritmasından çok daha hızlı olarak yapabilmektedir.

Tablo 5. BCOFI Algoritmasının bazı doğal verikümeleri için elde edilen 5-katlı çapraz-sağlama sınıflandırma doğruluğu ve programın Sparc 20/61 iş istasyonundaki ortalama çalışma süreleri. Bir karşılaştırma olması için aynı değerler 1-NN için de verilmiştir.

	BCOFI		1-NN	
Veri Kümesi	Doğruluk (%)	Süre (ms)	Doğruluk (%)	Süre (ms)
arrhythmia	62.16	1598.6	53.04	33330.0
bcancerw	94.85	32.4	95.00	5013.8
cleveland	80.18	15.8	80.52	693.0
dermatology	96.18	79.4	94.80	2735.2
diabetes	55.99	29.8	72.52	3731.2
glass	58.88	15.0	68.66	200.8
horse	78.81	22.0	78.24	1110.2
hungarian	84.64	15.0	75.50	719.6
ionosphere	81.18	71.6	84.62	1917.6
iris	96.00	4.4	93.98	190.2
liver	53.04	9.6	63.28	658.4
musk	79.19	414.0	73.10	17755.6
vehicle	56.75	82.0	67.22	7615.4
wine	96.65	17.6	94.40	342.0

Geliştirilen BCOFI algoritması VFI (Voting Feature Intervals) adıyla 9. Avrupa Makina Öğrenmesi konferansında bir bildiri ile sunulmuş ve bildiriler kitabında yayınlanmıştır. Bildirinin bir kopyası Ek-5'de verilmiştir. Ayrıca BCOFI algoritmasının dermatoloji verikümesine uygulaması 6. Türk Yapay Zeka ve Yapay Sinir Ağları Sempzyomunda bir bildiri ile sunulmuş ve "New Trends in Artificial Intelligence & Neural Networks" adlı bildiriler kitabında yayınlanmıştır. Bildirinin bir kopyası Ek-8'de sunulmuştur. BCOFI algoritmasının aritmi verikümesine uygulaması 24. Kardiyolojide Bilgisayarlar Konferansında sunulmuş ve daha sonra basılacak olan bildiriler kitabında yayınlanmak üzere kabul edilmiştir. Bildirinin bir kopyası Ek-10'de verilmiştir.

```
öğren (ÖğrenmeKümesi)
başlangıç
  her öznitelik f için
    eğer f lineer ise
      her sınıf c için
        /* ÖğrenmeKümesi içindeki örnekleri kullanarak */
        c sınıfındaki örneklerin f özniteliğindeki
                en küçük ve en büyük değerlerini uç_noktalar_dizisine ekle
      uç_noktalar_dizisini sırala
      /* her iki uç nokta arası bir çakışan öznitelik bölüntüsüdür */
    eğer f nominal ise
      her farklı değeri uç_noktalar_dizisine ekle
      /* her uç nokta bir çakışan öznitelik bölüntüsüdür */
  her öznitelik bölüntüsü b için
    her sınıf c için
      b bölüntüsündeki sınıfı c olan örnek sayısını bul
      b bölüntüsünde oy_c = (bölüntüdeki sınıfı c olan örnek sayısı) / (sınıfı c olan toplam örnek sayısı)
      bölüntünün toplam oyu 1 olacak şekilde sınıf oylarını normalize et
bitiş

sınıflandır (örnek, AğırlıkVektörü)
başlangıç
  her sınıfın oyunu 0'a eşitle
  her öznitelik f için
    /* örneğin f özniteliğindeki değerine karşı gelen bölüntüyü bul */
    bölüntü = bölüntüAra(f, örnek_f )
    her sınıf c için
      c'nin oyunu (bölüntü'nün c'ye verdiği oy × AğırlıkVektörü[f]) kadar artır
  tahmin = oyu en çok olan sınıf
  tahmini döndür
bitiş
```

Şekil 11. BCFI Algoritması.

Bir sınflandırma öğrenen bir programdan beklenen diğer bir özellik de öğrenilen sınıflandırma bilgisinin uzmanlar tarafından incelenebilecek bir yapıda olmasıdır. Bu nedenle BCOFI algoritması tarafından öğrenilen sınıflandırma bilgisinin uzman tarafından değerlendirilebilmesi ek bir çıktı programı hazırladık. Bu program her öznitelik için oluşturulan çakışan öznitelik aralıklarını ve sınıflandırma bilgilerini yazıcıda basılabilir bir formatta (ASCII) üretmektedir. Dermatoloji verikümesinin bazı öznitelikleri için üretilen çıktı Şekil 12'de sunulmuştur.

Feature 1

VOTES FOR CLASSES:	[1]	[2]	[3]	[4]	[5]	[6]
If value < 0.00 then:	(-)	(-)	(-)	(-)	(-)	(-)
If (*) value = 0.00 then:	0.15	(-)	0.23	(-)	0.63(+)	(-)
If 0.00 < value < 1.00 then:	(-)	(-)	(-)	(-)	(-)	(-)
If (*) value = 1.00 then:	0.06(-)	0.08(-)	0.09	0.21	0.46(+)	0.10
If 1.00 < value < 3.00 then:	0.16	0.15	0.19	0.18	0.12	0.21
If (*) value = 3.00 then:	0.29	0.29	0.16	0.10	0.05(-)	0.12
If 3.00 < value < 0.00 then:	(-)	(-)	(-)	(-)	(-)	(-)

Feature 6

VOTES FOR CLASSES:	[1]	[2]	[3]	[4]	[5]	[6]
If value < 0.00 then:	(-)	(-)	(-)	(-)	(-)	(-)
If (*) value = 0.00 then:	0.20	0.20	0.01(-)	0.20	0.20	0.20
If 0.00 < value < 3.00 then:	(-)	(-)	1.00(+)	(-)	(-)	(-)
If (*) value = 3.00 then:	(-)	(-)	1.00(+)	(-)	(-)	(-)
If 3.00 < value < 0.00 then:	(-)	(-)	(-)	(-)	(-)	(-)

Feature 15

VOTES FOR CLASSES:	[1]	[2]	[3]	[4]	[5]	[6]
If value < 0.00 then:	(-)	(-)	(-)	(-)	(-)	(-)
If (*) value = 0.00 then:	0.20	0.20	0.20	0.20	(-)	0.20
If 0.00 < value < 1.00 then:	(-)	(-)	(-)	(-)	(-)	(-)
If (*) value = 1.00 then:	(-)	(-)	(-)	(-)	1.00(+)	(-)
If 1.00 < value < 2.00 then:	(-)	(-)	(-)	(-)	(-)	(-)
If (*) value = 2.00 then:	(-)	(-)	0.06(-)	(-)	0.94(+)	(-)
If 2.00 < value < 3.00 then:	(-)	(-)	(-)	(-)	(-)	(-)
If (*) value = 3.00 then:	(-)	(-)	(-)	(-)	1.00(+)	(-)
If 3.00 < value < 0.00 then:	(-)	(-)	(-)	(-)	(-)	(-)

Feature 20

VOTES FOR CLASSES:	[1]	[2]	[3]	[4]	[5]	[6]
If value < 0.00 then:	(-)	(-)	(-)	(-)	(-)	(-)
If (*) value = 0.00 then:	0.01(-)	0.21	0.21	0.21	0.19	0.18
If 0.00 < value < 1.00 then:	(-)	(-)	(-)	(-)	(-)	(-)
If (*) value = 1.00 then:	0.49	(-)	(-)	(-)	0.14	0.37
If 1.00 < value < 2.00 then:	(-)	(-)	(-)	(-)	(-)	(-)
If (*) value = 2.00 then:	0.97(+)	(-)	(-)	(-)	0.03(-)	(-)
If 2.00 < value < 3.00 then:	(-)	(-)	(-)	(-)	(-)	(-)
If (*) value = 3.00 then:	1.00(+)	(-)	(-)	(-)	(-)	(-)
If 3.00 < value < 0.00 then:	(-)	(-)	(-)	(-)	(-)	(-)

Şekil 12. BCOFI tarafından dermatoloji verikümesi için üretilen çıktının bir bölümü.

Burada her satır ilgili özniteliğin bir değer aralığına karşı düşmektedir. Sütunlar ise sınıfları göstermektedir. Matrisin her bir değeri ise ilgili özniteliğin satırdaki değeri için sütundaki sınıfa verdiği oy miktarını belirtmektedir. Eğer bu oy tamamen sıfır ise

(-) işareti ile gösterilmektedir. Eğer ilgili özniteliğin satırdaki değer aralığı tek bir sınıf için belirleyici ise bu durum oy miktarnın yanında bir (+) ile gösterilmektedir.

Dermatoloji verikümesi için üretilen bu çıktı verikümesini hazırlayan uzman Doç. Dr. Nilsel İlter'e değerlendirmesi için sunulmuştur. Kendisi tüm çıktıyı detaylı olarak incelemiş, ve kendisinin hastalara tanı koyarken kullandığı bilginin bir kağıt üzerine dökülmüş hali ie karşılaştığını söylemiştir. Bunun ötesinde uzman bu bilginin dermatoloji konusunda uzmanlık eğitimi gören öğrencilere bu bilgilerin öğretilebileceğini belirtmiştir. Bunun üzerine şu anda bir lisans öğrencisi öğrenilen bu bilgileri ve bu bilgileri kullanarak sınıflandırma yapabilecek ve Gazi Üniversitesi Tıp Fakültesinde kullanılmakta olan PC türü bilgisayarlarda çalışabilecek bir program hazırlamaktadır.

6 ÖZNİTELİK AĞIRLIKLARININ GENETİK ALGORİTMA İLE ÖĞRENİLMESİ

BFCP ve BCOFI algoritmaları sınıflandırma sırasında dışarıdan verilen öznitelik ağırlıklarını kullanabilmektedirler. Bu öznitelik ağırlıkları doğru olarak verildiğinde hen BCFP algoritmaları ve hemde BCOFI algoritması daha yüksek bir doğrulukla sınıflandırma yapabileceklerdir. Bu kapsamda genetik algoritmanın amacı verilen bir sınıflandırıcının yine verilen bir veri kümesi üzerinde doğruluğunu maksimize eden öznitelik ağırlıklarının öğrenilmesidir. Bu proje kapsamında geliştirdiğimiz BCFP ve BCOFI algoritmaları için ayrı ayrı genetik algoritmalar yerine herhangi bir sınıflandırıcı ile kullanılabilecek tek ve genel bir genetik algoritma geliştirilmiştir. Bu genetik algoritmanın genel yapısı Şekil 13'de verilmiştir.

Geliştirilen genetik algoritmada her bir kromozom veri kümesindeki öznitelik sayısı kadar gen içerir. Her bir gen karşı düşen özniteliğin ağırlığını temsil etmektedir (bu değer 0 ile 1 arasında reel bir sayıdır). Öznitelik ağırlıklarının (gen değerleri) toplamı daima 1'dir Böyelikle herhangi bir özniteliğin toplam içindeki payı diğer öznitelik ağırlıklarından bağımsız olarak belirlenebilmektedir.

Başlangıç populasyonu, gen değerleri toplamı 1 olacak şekilde, rasgele seçilmiş sayılardan oluşan kromozomlardan oluşturulur. Bu projede gerçekleştirilen deneylerde populasyon büyüklüğü 100 kromozom olacak şekilde belirlenmiştir. Genetik algoritma her biri aşağıdaki adımlardan oluşan en çok 100 nesil üretecek şekilde çalıştırılmıştır:

1. Yeni populasyon üretimi (reproduction)
2. Çaprazlama (crossover)
3. Mutasyon (mutation)
4. Elitizm (elitism)

Yeni populasyon üretiminde rulet-tekeri yöntemi kullanılmıştır. Çaprazlama işlemi olarak *sürekli üniform çaprazlama* (*continuous uniform crossover*) adı verilen yeni bir çaprazlama operatörü tanımlanmıştır. Buna göre rasgele seçilen iki kromozomdan iki yeni kromozom üretilir. Bu işlem çaprazlama olasılığı (p_c) ile

uygulanır. Eğer çaprazlama uygulanmayacak ise (1 - p_c olasıkla) eski kromozomlar yeni nesile aynen aktarılır. Buna göre,

$\boldsymbol{x} = < x_1, x_2, \ldots, x_n >$ ve $\boldsymbol{y} = < y_1, y_2, \ldots, y_n >$

şeklinde iki kromozom verildiğinde oluşturulacak yeni kromozomlar

$\boldsymbol{x}' = < x_1', x_2', \ldots, x_n' >$ ve $\boldsymbol{y}' = < y_1', y_2', \ldots, y_n' >$ de

$x_i' = s \times x_i + (1-s) \times y_i$ ve

$y_i' = s \times y_i + (1-s) \times x_i$

şeklinde belirlenir. Burada *s* çaprazlamanın *adım* (*stride*) adı verilen bir parametresidir. Kolaylıkla kanıtlanabilir ki, *s* nin bütün değerleri için, eğer çaprazlamadan önce $\sum_{i=1}^{n} x_i = 1$ ve $\sum_{i=1}^{n} y_i = 1$ ise çaprazlamadan sonra $\sum_{i=1}^{n} x_i' = 1$ ve $\sum_{i=1}^{n} y_i' = 1$ dir. Bu çaprazlama işlemine sürekli üniform çaprazlama (continuous uniform crossover) adı verilmiştir.

Mutasyon işlemi ise, ardışık iki genin mutasyon olasılığı (p_m) ile yerlerinin değiştirilmesi şeklinde yapılır. Böylelikle bir kromozomun gen toplamının 1 etme şartı bozulmamış olur.

Son olarak populasyonda bulunan en iyi kromozom bir sonraki nesile aktarılmak üzere işaretlenir. Genetik algoritma her bir adımda bu işlemleri uygulayarak yeni nesiller üretmeye devam eder. Bu projede ençok 100 nesil üretilmesine izin verilmiştir.

Yeni bir neslin üretilmesinde kullanılan ilk işlem olan yeni_populasyon_üret programı uygunluğu (fitness) yüksek olan kromozomları daha yüksek bir olasılıkla seçmektedir. Bir kromozomun uygunluğu olarak, bu projede, kromozomun gen değerlerinin öznitelik ağırlıkları olarak alınmasıyla kullanılan sınıflandırıcının veri kümesi üzerindeki 5-katlı çapraz sağlama doğruluğu alınmıştır. Bu nedenle

bcfp (*ÖğrenmeKümesi*, *TestKümesi*, *AğırlıkVektörü*) ve

bcofi (*ÖğrenmeKümesi*, *TestKümesi*, *AğırlıkVektörü*)

```
genetik_algoritma ()
başlangıç
   adım = 0
   her kromozomun genleri toplamı 1 olacak şekilde ilk populasyonu (P(adım)) oluştur
      /* kromozomdaki her bir gen karşı düşen özniteliğin ağırlığını gösterir */
      değerle(P(adım))
      bitme şartı sağlanıncaya kadar
         adım = adım + 1
         yeni_populasyon_üret(P(adım))
         çaprazla(P(adım))
      mutasyon_uygula(P(adım))
      elitizm_uygula(P(adım))
      değerle(P(adım))
   yap
bitiş

yeni_populasyon_üret (Populasyon)
başlangıç
   toplam-uygunluk = Populasyon'daki kromozomların uygunluklarının toplamı
   KromozomSayısı kadar tekrarla
      eski Populasyondan rasgele bir kromozomu yeni Populasyona kopyala
      /* k kromozomunun seçilme olasılığı = k'nın uygunluğu / toplam uygunluk */
bitiş

çaprazlama (Populasyon) /* 2-noktalı çaprazlama */
başlangıç
   Populasyon 'daki her iki kromozom çifti k ve k+1 için
      çaprazlama olasılığı ile    /* çaprazlama olasılığı, p_c = 0.8 */
         çapraz-nokta1 = [2..GenSayısı) aralığında rasgele bir sayı
         çapraz-nokta2 = (çapraz-nokta1 ..GenSayısı] aralığında rasgele bir sayı
         çapraz-nokta1 ve çapraz-nokta2 arasındaki genlerin yerlerini değiştir
         yeni k ve k+1 kromozomlarını normalize et
bitiş

mutasyon (Populasyon)
başlangıç
   Populasyon 'daki her kromozom k için
      mutasyon olasılığı ile    /* mutasyon olasılığı, p_m = 0.001 */
         kromozom k'daki her gen g için
            g genini g +1 geni ile değiştir
bitiş
```

Şekil 13. Genetik Algoritma.

adlı fonksiyonlar tanımlanmıştır. Bu fonksiyonların değeri ilgili sınıflandıcının *ÖğrenmeKümesi*, ile öğretilmesinden sonra *AğırlıkVektörü* kullanılarak *TestKümesi* üzerindeki ortalama sınıflandırma doğruluğudur. Bir populasyondaki tüm kromozomların uygunluk değerlerinin belirlenmesinde kullanılan *değerle* programı Şekil 14'de verilmiştir.

BCFP ve BCOFI sınıflandırıcılarında öznitelik ağırlıklarının eşit alınması yerine genetik algoritma ile öğrenilen ağırlıkların kullanılması ile elde edilen doğruluklar karşılaştırmalı olarak Tablo 6'da verilmiştir. Buna göre genetik algoritma her iki sınıflandırıcı için tüm veri kümelerinde önemli doğruluk artışları sağlamıştır. Ancak genetik algoritmanın çalışma süresi beklendiği gibi oldukça uzundur. Örneğin BCOFI sınıflandırıcısı ve aritmi veri kümesi için 1 neslin üretilmesi (ortalama 40 çaprazlama, 13 mütasyon ve 85 uygunluk değerlendirmesi) yaklaşık yarım saat sürmektedir.

```
Değerle (Populasyon)
başlangıç
   Populasyon'daki her kromozom için
      /* kromozom öznitelik ağırlıkları vektörüdür */
      doğruluk = 5-katlı-çapraz-sağlama(Sınıflandırıcı, VeriKümesi, AğırlıkVektörü)
      doğruluk'u döndür  /* doğruluk kromozomun uygunluğudur */
bitiş

5-katlı-çapraz-sağlama (Sınıflandırıcı, VeriKümesi, AğırlıkVektörü)
başlangıç
   VeriKümesi'ni 5 eşit parçaya böl
   5 defa tekrarla
      TestKümesi = test için henüz hiç kullanılmamış bir parça
      ÖğrenmeKümesi = VeriKümesi - TestKümesi
      doğruluk = Sınıflandırıcı (ÖğrenmeKümesi, TestKümesi, AğırlıkVektörü)
   5 doğruluk'un ortalamasını gönder
bitiş
```

Şekil 14. Bir kromozonun uygunluk değerinin belirlenmesi.

Bu tablodan da görüleceği üzere genetik algoritma yüksek doğrulıkta sınıflandırma yapılabilecek şekilde öznitelik ağırlıklarını doğru olarak belirlemiştir. Buradaki tek problem genetik algoritmanın çalışma süresinin çok uzun olmasıdır. Ancak, öğrenme aşamasının sadece bir uygulandığı gözönünde tutulursa, yüksek doğruluk için buna katlanmak mümkündür.

Tablo 6. BCFP.4 ve BCOFI Algoritmalarının bazı doğal verikümeleri için elde edilen 5-katlı çapraz-sağlama sınıflandırma doğrulukları. Burada her bir sınıflandırıcı için önce öznitelik ağırlıklarının eşit alındığı durumdaki doğrulukları sonrada genetik algoritma ile öğrenilen ağırlıklar kullanılarak elde edilen doğruluklar verilmiştir. Ayrıca genetik algoritmanın bir adım ilerlemesi için SUN-Sparc 20/61 iş istasyonunda gereken süre verilmiştir.

	BCFP.4			BCOFI		
	Eşit ağırlık	Genetik algoritma		Eşit ağırlık	Genetik algoritma	
Veri kümesi	Doğruluk (%)	Doğruluk (%)	Süre (dk:sn)	Doğruluk (%)	Doğruluk (%)	Süre (dk:sn)
arrhythmia	57.72	59.04	44:57	62.16	66.81	30:24
bcancerw	97.00	97.86	6:03	94.85	96.28	6:29
cleveland	80.50	86.46	4:55	80.18	84.81	6:30
dermatology	58.50	70.50	7:38	96.18	98.36	7:06
diabetes	69.76	72:72	12:17	55.99	66.53	7:05
glass	45.36	60.28	4:13	58.88	63:58	5:43
horse	76.36	84.78	6:40	78.81	85.59	6:01
hungarian	75.48	84.64	5:47	84.64	86.68	8:02
ionosphere	88.88	91:16	7:44	81.18	84.61	8:35
iris	90.66	96.02	3:35	96.00	96.67	5:12
liver	59.72	66.38	4:33	53.04	61.74	5:44
musk	73.34	77.34	28:06	79.19	82.35	20:01
vehicle	58.14	62.64	8:45	56.75	59.70	8:13
wine	89.92	93.84	4:49	96.65	99.43	6:01

Örnek olarak dermatoloji verikümesi için BCOFI algoritmasıyla öğrenilen öznitelik ağırlıkları aşağıda verilmiştir:

$w_1 = 0.01$, $w_2 = 0.0154$, $w_3 = 0.0361$, $w_4 = 0.0259$, $w_5 = 0.0499$,
$w_6 = 0.0458$, $w_7 = 0.0317$, $w_8 = 0.0362$, $w_9 = 0.0191$, $w_{10} = 0.0148$,
$w_{11} = 0.0393$, $w_{12} = 0.0215$, $w_{13} = 0.027$, $w_{14} = 0.0205$, $w_{15} = 0.0368$,
$w_{16} = 0.0197$, $w_{17} = 0.0194$, $w_{18} = 0.0288$, $w_{19} = 0.0274$, $w_{20} = 0.024$,
$w_{21} = 0.0314$, $w_{22} = 0.0191$, $w_{23} = 0.0316$, $w_{24} = 0.0295$, $w_{25} = 0.033$,
$w_{26} = 0.0577$, $w_{27} = 0.032$, $w_{28} = 0.0318$, $w_{29} = 0.0452$, $w_{30} = 0.0373$,
$w_{31} = 0.0264$, $w_{32} = 0.0301$, $w_{33} = 0.0348$, $w_{34} = 0.0107$

Buna göre f_{26} (disappearence of granular layer) en çok önemli (ayırıcı) f_1 (erythema) ise en az önemli öznitelik olarak belirlenmektedir. Bu bulgular verileri hazırlayan uzman ile incelenmiş ve kendilerinin (sayısal olmayan) değerlendirmeleri ile çok uyumlu olduğu saptanmıştır. Örneğin, uzman erythema özniteliğinin bu tür hastalıkların hepsinin genel özelliği olduğunu dolayısıyla etkisinin az olacağını, aynı şekilde hasta yaşının çok önemli bir veri olmadığını ancak C_6 (pityriasis rubra pilaris) hastalığının genellikle çocuklarda görüldüğünü belirmiştir. Bunun yanında f_{26} özniteliğinin C_1 hastalığını, f_5, f_6 ve f_{33} özniteliklerinin C_3 ve C_4 hastalıklarını, f_{21} özniteliğinin C_1 ve C_5 hastalıklarını, f_{28} özniteliğinin C_2 hastalığını, f_{29} özniteliğinin C_3

hastalığını ve f_{30} özniteliğinin C_6 hastalığını belirlemede önemli olduğunu kabul etmiştir.

Geliştirilen bu genetik algoritma önce *k*-NN sınıfladırıcısında öznitelik ağırlıklarının öğrenilmesi için kullanılmış ve sonuçlar 6. Belçika-Hollanda Makina Öğrenmesi konferansında sunulmuş ve bildiri kitabında yayınlanmıştır. Bildirinin bir kopyası Ek-6'da verilmiştir. Daha sonra aynı genetik algoritma *k*-NNFP sınıfladırıcısında öznitelik ağırlıklarının öğrenilmesi için kullanılmış ve sonuçlar 12. Uluslararası Bilgisayar ve Enformatik Bilimleri Sempozyumunda sunulmuş ve bildiri kitabında yayınlanmıştır. Bildirinin bir kopyası Ek-8'da verilmiştir.

7 SONUÇ

Bu araştırma projesinde sınıflandırmaya yönelik olarak kavramların öznitelik bölümlemesi ve çakışan öznitelik aralıkları ve toplu bir şekilde ve mekanik olarak öğrenilebilmesi ve ayrıca öznitelik ağırlıklarının genetik algoritmalar yardımıyla öğrenilebilmesi için çeşitli algoritma ve teknikler geliştirilmiştir. Ayrıca, geliştirilen tekniklerin gerçek hayat problemlerine bir uygulamasının oluşturulması ve tekniklerin sandart verikümelerinin dışındaki alanlarda da sınanabilmesi için tıbbın iki alanında birer veri kümesi de oluşturulmuştur. Dermatoloji ve Aritmi adları verilen bu veri kümeleri, Irvine'daki Kalifornia Üniversitesinde tutulmakta olan UCI-Repository adındaki listeye alınmıştır. Bu sayede makina öğrenmesi konusunda geliştirilecek olan algoritmaların değerlendirilmesinde ve karşılaştırılmasında deneme kümesi olarak kullanılacaktır. Bu şekilde de projenin uluslararası düzeyde tanıtımı sağlanmış olunmaktadır.

Amaçlanan hedef verilen bir sınıflandırılmış örnek verikümesinden, o verikümesinin özelliklerini öğrenerek, yeni gelecek örneklerin hızlı ve yüksek doğrulukla sınıflandırılmasını yapabilecek bir sistemin geliştirilmesiydi. Makina öğrenmesi alanında bugüne kadar geliştirilen sistemlerden hızlı sınıflandırma yapanlar yüksek doğruluk elde edememekte, buna karşılık yüksek doğrulukla sınıflandırma yapabilenler ise çok yavaş çalışmaktadırlar. Bu projede doğruluk derecesinden ödün vermeden hızlı öğrenen ve hızlı sınıflandırma yapabilmeye yönelik olarak geliştirilen *öznitelik bölümlemesi* ve *çakışan öznitelik aralıkları* bilgi gösterim tekniklerinin başarılı oldukları görülmüştür.

Bunun ötesinde, geliştirilen her iki sınıflandırma algoritması verilen öznitelik ağırlıklarını da kullanabilmektedir. Bu nedenle, proje önerisinde amaçladığı şekilde, öznitelik ağırlıklarının da mekanik olarak öğrenilebilmesi için genetik algoritmalardan yararlanıldı. Bu şekilde öğrenilen öznitelik ağırlıklarını da kullanarak, geliştirilen sınıflandırma programlarının daha yüksek doğruluk dereceleri elde ettikleri gözlendi. Genetik algoritmalarının çalışma süresi genel olarak oldukça uzundur, ancak her verikümesi için sadece bir kez çalıştırılacağı göz önünde tutulursa, yüksek doğrulukta sınıflandırma yapabilmek için buna katlanılabilir.

Bu araştırmalar göstermiştir ki geliştirilen bilgi gösterim teknikleri ve ilgili makina öğrenmesi ve sınıflandırma teknikleri, tıp gibi, matematiksel medellerin bulunmadığı, ancak çok sayıda çözüm veya sınıflama örneğinin bulunuduğu problemlerin çözümünde başarı ile uygulanabilecektir (Güvenir, 1997). Geliştirilen algoritmaların tıpdaki uygulamalarının özetlendiği ve Popüler Bilim adlı dergide yayınlanan makalenin bir kopyası Ek-9'da verilmiştir. Bu sınıflandırma öğrenme algoritmasının kullanılabilmesi için örnekler, öncelikle öznitelik değerlerinden ve ait oldukları sınıf değerlerinden oluşan birer verktör ile gösterilebilmelidir.

Burada geliştirilen teknikler kullanılarak daha önce oluşturulmuş veritabanlarından da yararlanılabilir. Örneğin, bankalar tarafından oluşturulan veritabanlarındaki bilgiler kullanılarak, hangi kredi başvurularının sorun çıkarabileği belirlenebilir. Diğer taraftan üretim tesislerinde daha önceki deneyimler kullanılarak verilen yeni bir durum için en optimum parametre değerleri belirlenebilir.

İleriye dönük araştırma projesi olarak bu proje çerçevesinde geliştirilmiş olan öğrenme tekniklerin veri madenciliği (data mining) konusuna uygulanabilirliği araştırılabilir. Veri madenciliği alanında makina öğrenmesi teknikleri üç temel şekilde kullanılabilmektedir. Bunlar, sınıflandırma (classification) bilgisinin öğrenilmesi, ilişkilendirme kurallarının (association rules) öğrenilmesi ve sıra (sequence) öğrenimi olarak tanımlanmaktadır (Agrawal, Imielinski, Swami, 1993). Veritabanlarında uygulanacak sınıflandırma algoritmalarından beklenen en önemli özellik, veritabanlarının büyüklüğü nedeniyle, hızlı öğrenme ve aynı zamanda hızlı sınıflandırma yapabilmeleridir. Daha önce de gösterildiği üzere bu proje çerçevesinde geliştirilmiş olan sınıflandırma öğrenme algoritmaları bu iki şartı kolaylıkla sağlamaktadırlar. Veri madenciği uygulamalarında aranan diğer bir özellik de öğrenilen sınıflandırma bigisinin uzmanlar tarafından anlaşılıp değerlendirilebilecek bir şekilde sunulabilmesidir. Proje kapsamında geliştirilen programlar bu şatı da sağlamaktadırlar. İşte bu nedenlerden ötürü geliştirilen sınıflandırma öğrenmesi tekniklerinin veri madenciliğinde uyglama alanları araştırılmalıdır.

Kaynakça

Agrawal R., Imielinski T, Swami A., Database Mining: A Performance Perspective, *IEEE Transactions on Data Engineering, Special Issue on Learning and Discovery in Knowledge-Based Databases*, 5, 914 - 925, (1993).

Aha D. W., Kibler D., Albert M. K., Instance-Based Learning Algorithms, Machine Learning, 6, 37-66, (1991).

Aha D. W., Incremental, Instance-Based Learning of Independent and Graded Concept Descriptions, *Proceedings of the Sixth International Workshop on Machine Learning*, Morgan Kaufmann, (1989), pp: 387-391.

Akkuş A., *Batch Learning of Disjoint Feature Intervals*, (Yüksek Lisans Tezi), Bilkent Üniversitesi, Bilgisayar ve Enformatik Mühendisliği Bölümü, (1996).

Akkuş A., Güvenir H. A., K Nearest Neighbor Classification on Feature Projections, *13th Int. Conference on Machine Learning*, (*ICML'96*), ed: Saitta L., Morgan Kaufmann, Bari, Italy, (1996), pp: 12-19.

De Jong K., Learning with Genetic Algorithms: An Overwiew, *Machine Learning*, 3, 121-138 (1988).

Demiröz G., *Non-incremental Classification Learning Algorithms Based on Voting Feature Intervals*, (Yüksek Lisans Tezi), Bilkent Üniversitesi, Bilgisayar ve Enformatik Mühendisliği Bölümü, (1997).

Demiröz G., Güvenir H. A., İlter N., Differential Diagnosis of Eryhemato-Squamous Diseases using Voting Feature Intervals, *6th Turkish Symposium on Artificial Intelligence and Neural Networks* (*TAINN'97*), Kızılcahamam, Ankara, (1997), pp: 190-194.

Demiröz G., Güvenir H. A., Classification by Voting Feature Intervals, *9th European Conference on Machine Learning* (*ECML-97*), ed: van Someren M. and Widmer G., Springer-Verlag, *LNAI 1224*, Prague, Czech Republic, (1997), pp: 85-92.

Demiröz G., Güvenir H. A., Genetic Algorithms to Learn Feature Weights for the Nearest Neighbor Algorithm*, 6th Belgian-Dutch Conference on Machine Learning*

(*BENELEARN-96*), ed: H. J. van den Herik and T. Weijters, Universiteit Maastricht, The Netherlands, (1996), pp: 117-126.

Duda R.D., Hart P.E., *Pattern Classification and Scene Analysis*, Wiley, New York, (1973).

Goldberg D.E., Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, (1989).

Güvenir H. A., Makina Öğrenmesi Tıbbın Hizmetinde, Popüler Bilim. Sayı: 46 (1997), pp: 25-27.

Güvenir H. A., Akkuş A., Weighted K Nearest Neighbor Classification on Feature Projections, *12th Int. Symposium on Computer and Information Sciences*, Antalya, (1997), pp: 44-51.

Güvenir H. A., Acar B., Demiröz G., Çekin A., A Supervised Machine Learning Algorithm for Arrhythmia Analysis, *24th Computers in Cardiology Conference*, Lund, Sweden, (1997).

Güvenir H. A., Şirin İ., Classification by Feature Partitioning, *Machine Learning*, 23, 47-67, (1996).

Güvenir H. A., Şirin İ., Complexity of the CFP, a Method for Classification Based on Feature Partitioning*, Third Congress of the Italian Association for Artificial Intelligence*, Springer-Verlag *Advances in Artificial Intelligence*, *LNAI 728*, (1993) pp: 202-207.

Güvenir H. A., Şirin İ., A Genetic Algorithm for Classification by Feature Partitioning, *Fifth International Conference on Genetic Algorithms*, (1993) pp: 543-548.

Holte R. C., Very Simple Classification Rules Perform well on most Commonly used Datasets, *Machine Learning*, 23, 63-91, (1993).

Kelly J. D., Davis L., A Hybrid Genetic Algorithm for Classification *Twelfth International Joint Conference on Artificial Intelligence*, (1991) pp: 645-650.

Langley P., *Elements of Machine Learning*, Morgan Kaufmann, San Francisco, (1996).

Medin D.L., Schaffer M.M., Context Theory of Classification Learning, *Psychological Review*, 85, 207-238 (1978).

Mitchell T., *Machine Learning*, McGraw-Hill, New York, (1997).

Quinlan J. R., Induction of Tecision Trees, Machine Learning, 1, 81-106, (1986).

Salzberg S., A Nearest Hyperrectangle Learning Method, *Machine Learning*, 6, 251-276, (1991).

Salzberg S., *Learning with Generalized Examplars*, Kluwer Academic Publishers, Massachusettes (1990).

Spackman A. K., Learning Categorical Decision Criteria in Biomedical Domains, *5th Int. Conference on Machine Learning*, Ann Arbor, U.S.A. (1988) pp:36-46.

Şirin İ., Güvenir H. A., Empirical Evaluation of the CFP Algorithm, *Sixth Australian Joint Conference on Artificial Intelligence*, Melbourne, Australia (1993) pp: 311-315.

Ünsal-Koç H., *Classification with Overlapping Feature Intervals*, (Yüksek Lisans Tezi), Bilkent Üniversitesi, Bilgisayar ve Enformatik Mühendisliği Bölümü, (1995).

Wettschereck D., Aha D. W., Weighting Features, 1st Int. Conference on Case-Based Reasoning, Sesimbra, Portugal (1995), pp: 347-358.

Ekler

Ek-1: Uygulama Özeti

Ek-2: Dermatology.names adlı dosya

Ek-3: Arrhythmia.names adlı dosya

Ek-4: Akkuş A., Güvenir H. A., K Nearest Neighbor Classification on Feature Projections, *13th Int. Conference on Machine Learning*, (*ICML'96*), ed: Saitta L., Morgan Kaufmann, Bari, Italy, (1996), pp: 12-19.

Ek-5: Demiröz G., Güvenir H. A., Classification by Voting Feature Intervals, *9th European Conference on Machine Learning* (*ECML-97*), ed: van Someren M. and Widmer G., Springer-Verlag, *LNAI 1224*, Prague, Czech Republic, (1997), pp: 85-92.

Ek-6: Demiröz G., Güvenir H. A., Genetic Algorithms to Learn Feature Weights for the Nearest Neighbor Algorithm*, 6th Belgian-Dutch Conference on Machine Learning* (*BENELEARN-96*), ed: H. J. van den Herik and T. Weijters, Universiteit Maastricht, The Netherlands, (1996), pp: 117-126.

EK-7: Güvenir H. A., Akkuş A. Weighted K Nearest Neighbor Classification on Feature Projections, *12th Int. Symposium on Computer and Information Sciences*, Antalya, (1997), pp: 44-51.

EK-8: Demiröz G., Güvenir H. A., İlter N., Differential Diagnosis of Eryhemato-Squamous Diseases using Voting Feature Intervals, *6th Turkish Symposium on Artificial Intelligence and Neural Networks* (*TAINN'97*), Kızılcahamam, Ankara, (1997), pp: 190-194.

Ek-9: Güvenir H. A., Makina Öğrenmesi Tıbbın Hizmetinde, *Popüler Bilim*, Sayı: 46 (1997), pp: 25-27.

Ek-10: Güvenir H. A., Acar B., Demiröz G., Çekin A., A Supervised Machine Learning Algorithm for Arrhythmia Analysis, *24th Computers in Cardiology Conference*, Lund, Sweden, (1997).

Ek-11: Güvenir H. A., Demiröz G., İlter N., Learning Differential Diagnosis of Erythemato-Squamous Diseases using Voting Feature Intervals, *Artificial Intelligence in Medicine* (basımda).

EK-1

Uygulama Özeti

Bu proje çerçevesinde geliştirilen sınıflandırma öğrenmesi algoritmaları Üniversitelerin tıp ve iktisat gibi fakültelerine duyurulabilir. Bu tür konularda çalışan araştırmacılar ellerinde çok sayıda örnek olduğu taktirde geliştirdiğimiz algoritmaları kullanarak otomatik sınıflandırma yapabilecekleri gibi örneklerin kodlanmasında kullanılan özniteliklerin bağıl ağırlıklarını da belirleyebilirler. Ayrıca, üretim yapan kuruluşlar, üretim parametrelerinin en optimum değerlerinin bulunmasında bu projede geliştirilen öğrenme tekniklerinden yararlanabilirler. Sigorta şirketleri veya kredi veren kuruluşlar müşterilerinin risk derecelerini belirlemede geliştirdiğimiz sınıflandırma algoritmalarından yararlanabilirler.

İlgilenen araştırmacılardan beklenen ellerindeki verilerin özelliklerini belirten bir info dosyası (Ek-2 ve Ek-3 deki örneklerde olduğu şekilde) ve her satırında bir örneğin kodlandığı data dosyası olmak özere, her bir verikümesi için iki dosya, hazırlamalarıdır. Data dosyasındaki satırlarda önce özniteli değerleri boşluk (space) ya da "," karakteri ile ayırılarak yazılmalı, satırdaki son değer örneğin sınıfını belirtmelidir. Bu şekilde hazırlanmış dosyalar guvenir@cs.bilkent.edu.tr adresine elektronik posta yoluyla gönderilebilir. İlgilenen kurum veya kuruluşlara, kendi durumlarını göz önünde tutarak bu tür verilerin hazırlanmasında ve sonuçların değerlendirilmesinde yardımcı olacağımız duyurulmalıdır.

Dermatology.names adlı dosya

```
1. Title: Dermatology Database

2. Source Information:
   (a) Creators:
       -- 1. Nilsel Ilter, M.D., Ph.D., Gazi University, Ankara, Turkey
       -- 2. H. Altay Guvenir, PhD., (guvenir@cs.bilkent.edu.tr)
             Bilkent University, Ankara, Turkey
   (b) Donor: H. Altay Guvenir, PhD., (guvenir@cs.bilkent.edu.tr)
              Bilkent University, Ankara, Turkey
   (c) Date: November, 1997

3. Past Usage:
   1. G. Demiroz, H. A. Govenir, and N. Ilter,
      "Differential Diagnosis of Eryhemato-Squamous Diseases using
       Voting Feature Intervals" in: New Trends in Artificial Intelligence
       and Neural Networks (Proceedings of TAINN'97) Turhan Ciftcibasi,
       Mustafa Karaman and Volkan Atalay (Eds.), Kizilcahamam, Ankara,
      (May 22-23, 1997), 190-194

4. Relevant Information:
     This database contains 34 attributes, 33 of which are linear
     valued and one of them is nominal.

     The differential diagnosis of erythemato-squamous diseases is a real
     problem in dermatology. They all share the clinical features of
     erythema and scaling, with very little differences. The diseases in
     this group are psoriasis, seboreic dermatitis, lichen planus,
     pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris.
     Usually a biopsy is necessary for the diagnosis but unfortunately
     these diseases share many histopathological features as
     well. Another difficulty for the differential diagnosis is that a
     disease may show the features of another disease at the beginning
     stage and may have the characteristic features at the following
     stages. Patients were first evaluated clinically with 12 features.
     Afterwards, skin samples were taken for the evaluation of 22
     histopathological features. The values of the histopathological
     features are determined by an analysis of the samples under a
     microscope.

     In the dataset constructed for this domain, the family history feature
     has the value 1 if any of these diseases has been observed in the
     family, and 0 otherwise. The age feature simply represents the age of
     the patient. Every other feature (clinical and histopathological) was
     given a degree in the range of 0 to 3. Here, 0 indicates that the
     feature was not present, 3 indicates the largest amount possible,
     and 1, 2 indicate the relative intermediate values.

     The names and id numbers of the patients were recently
     removed from the database.

5. Number of Instances: 366

6. Number of Attributes: 34

7. Attribute Information:
   Complete attribute documentation:
     Clinical Attributes: (take values 0, 1, 2, 3, unless otherwise stated)
      1: erythema
      2: scaling
      3: definite borders
```

4: itching
5: koebner phenomenon
6: polygonal papules
7: follicular papules
8: oral mucosal involvement
9: knee and elbow involvement
10: scalp involvement
11: family history, (0 or 1)
34: Age (linear)

Histopathological Attributes: (take values 0, 1, 2, 3)
12: melanin incontinence
13: eosinophils in the infiltrate
14: PNL infiltrate
15: fibrosis of the papillary dermis
16: exocytosis
17: acanthosis
18: hyperkeratosis
19: parakeratosis
20: clubbing of the rete ridges
21: elongation of the rete ridges
22: thinning of the suprapapillary epidermis
23: spongiform pustule
24: munro microabcess
25: focal hypergranulosis
26: disappearance of the granular layer
27: vacuolisation and damage of basal layer
28: spongiosis
29: saw-tooth appearance of retes
30: follicular horn plug
31: perifollicular parakeratosis
32: inflammatory monoluclear inflitrate
33: band-like infiltrate

8. Missing Attribute Values: Several. Distinguished with '?'.

9. Class Distribution:
Database: Dermatology

Class code:	Class:	Number of instances:
1	psoriasis	112
2	seboreic dermatitis	61
3	lichen planus	72
4	pityriasis rosea	49
5	cronic dermatitis	52
6	pityriasis rubra pilaris	20

EK-3

Arrhythmia.names adlı dosya

```
1. Title: Cardiac Arrhythmia Database

2. Source Information:
   (a) Creators:
       -- 1. Burak Acar, M.S.: Bilkent University,
             EE Eng. Dept. Ankara, Turkey
       -- 2. Haldun Muderrisoglu, M.D., Ph.D., Baskent University, Ankara
   (b) Donor: H. Altay Guvenir, PhD., (guvenir@cs.bilkent.edu.tr)
              Bilkent University, Ankara, Turkey
   (c) Date: November, 1997

3. Past Usage:
   1. H. Altay Guvenir, Burak Acar, Gulsen Demiroz, Ayhan Cekin
      "A Supervised Machine Learning Algorithm for Arrhythmia Analysis"
      Proceedings of the Computers in Cardiology Conference,
      Lund, Sweden, 1997.

4. Relevant Information:
     This database contains 279 attributes, 206 of which are linear
     valued and the rest are nominal.

     Concerning the study of H. Altay Guvenir: "The aim is to distinguish
     between the presence and absence of cardiac arrhythmia and to
     classify it in one of the 16 groups. Class 01 refers to 'normal'
     ECG classes 02 to 15 refers to different classes of arrhythmia
     and class 16 refers to the rest of unclassified ones. For the
     time being, there exists a computer program that makes such a
     classification. However there are differences between the
     cardiolog's and the programs classification. Taking the
     cardiolog's as a gold standard we aim to minimise this difference
     by means of machine learning tools."

     The names and id numbers of the patients were recently
     removed from the database.

5. Number of Instances: 452

6. Number of Attributes: 279

7. Attribute Information:
   -- Complete attribute documentation:
      1 Age: Age in years , linear
      2 Sex: Sex (0 = male; 1 = female) , nominal
      3 Height: Height in centimeters , linear
      4 Weight: Weight in kilograms , linear
      5 QRS duration: Average of QRS duration in msec., linear
      6 P-R interval: Average duration between onset of P and Q waves
        in msec., linear
      7 Q-T interval: Average duration between onset of Q and offset
        of T waves in msec., linear
      8 T interval: Average duration of T wave in msec., linear
      9 P interval: Average duration of P wave in msec., linear
     Vector angles in degrees on front plane of:, linear
     10 QRS
     11 T
     12 P
     13 QRST
     14 J

     15 Heart rate: Number of heart beats per minute ,linear
```

```
Of channel DI:
 Average width, in msec., of: linear
 16 Q wave
 17 R wave
 18 S wave
 19 R' wave, small peak just after R
 20 S' wave

 21 Number of intrinsic deflections, linear

 22 Existence of ragged R wave, nominal
 23 Existence of diphasic derivation of R wave, nominal
 24 Existence of ragged P wave, nominal
 25 Existence of diphasic derivation of P wave, nominal
 26 Existence of ragged T wave, nominal
 27 Existence of diphasic derivation of T wave, nominal

Of channel DII:
 28 .. 39 (similar to 16 .. 27 of channel DI)
Of channels DIII:
 40 .. 51
Of channel AVR:
 52 .. 63
Of channel AVL:
 64 .. 75
Of channel AVF:
 76 .. 87
Of channel V1:
 88 .. 99
Of channel V2:
 100 .. 111
Of channel V3:
 112 .. 123
Of channel V4:
 124 .. 135
Of channel V5:
 136 .. 147
Of channel V6:
 148 .. 159

Of channel DI:
 Amplitude , * 0.1 milivolt, of
 160 JJ wave, linear
 161 Q wave, linear
 162 R wave, linear
 163 S wave, linear
 164 R' wave, linear
 165 S' wave, linear
 166 P wave, linear
 167 T wave, linear

 168 QRSA , Sum of areas of all segments divided by 10,
     ( Area= width * height / 2 ), linear
 169 QRSTA = QRSA + 0.5 * width of T wave * 0.1 * height of T
     wave. (If T is diphasic then the bigger segment is
     considered), linear

Of channel DII:
 170 .. 179
Of channel DIII:
 180 .. 189
Of channel AVR:
 190 .. 199
Of channel AVL:
```

```
      200 .. 209
     Of channel AVF:
      210 .. 219
     Of channel V1:
      220 .. 229
     Of channel V2:
      230 .. 239
     Of channel V3:
      240 .. 249
     Of channel V4:
      250 .. 259
     Of channel V5:
      260 .. 269
     Of channel V6:
      270 .. 279

8. Missing Attribute Values: Several.  Distinguished with '?'.

9. Class Distribution:
       Database:  Arrhythmia

       Class code :   Class   :                        Number of instances:
       01             Normal                                          245
       02             Ischemic changes (Coronary Artery Disease)       44
       03             Old Anterior Myocardial Infarction               15
       04             Old Inferior Myocardial Infarction               15
       05             Sinus tachycardy                                 13
       06             Sinus bradycardy                                 25
       07             Ventricular Premature Contraction (PVC)           3
       08             Supraventricular Premature Contraction            2
       09             Left bundle branch block                          9
       10             Right bundle branch block                        50
       11             1. degree AtrioVentricular block                  0
       12             2. degree AV block                                0
       13             3. degree AV block                                0
       14             Left ventricule hypertrophy                       4
       15             Atrial Fibrillation or Flutter                    5
       16             Others                                           22
```

BİBLİYOGRAFİK BİLGİ FORMU

1- Proje No: EEEAG-153 2- Rapor Tarihi: Eylül 1997

3- Projenin Başlangıç ve Bitiş Tarihleri: 1 Ağustos 1995 – 1 Ağustos 1997

4- Projenin Adı: ÖZNİTELİK BÖLÜMLEMESİ VE ÇAKIŞAN ÖZNİTELİK ARALIKLARI ŞEKLİNDE ÇOKLU KAVRAM TANIMLARI İÇİN TOPLU MAKİNA ÖĞRENMESİ TEKNİKLERİNİN GELİŞTİRİLMESİ

5- Proje Yürütücüsü ve Yardımcı Araştırıcılar: Doç. Dr. H. Altay Güvenir
Aynur Akkuş, Gülşen Demiröz

6- Projenin Yürütüldüğü Kuruluş ve Adresi: Bilkent Üniversitesi
Bilgisayar ve Enformatik Mühendisliği Bölümü
06533 Ankara

7- Destekleyen Kuruluş(ların) Adı ve Adresi:

8- Öz (Abstract):

Sınıflandırma problemleri çoklu kavram tanımlaması problemleri şeklinde modellenebilmektedir. Çoklu kavram tanımlarının sınıflandırılmış örneklerden öğrenilmesi makina öğrenmesi alanında yapılan çalışmalar arasında en çok uygulama alanı bulmuş olanlardan birisidir. Bu projede kavram tanımlarının hem öznitelik bölümlemesi ve hem de çakışan öznitelik aralıkları şeklinde toplu şekilde öğrenilmesi için algoritmalar geliştirildi. Bu çalışmada kavramlar her özniteliğin bir bölümlemesi veya her birisi bir kavrama karşı gelen çakışan öznitelik aralıkları şeklinde tanımlanmaktadır. Verilen yeni bir vakanın sınıflandırılması ise ayrı ayrı özniteliklere göre gerçekleştirilen sınıflandırılmalar arasında yapılan bir ağırlıklı oylama sonucunda belirlenmektedir. Özniteliklerin ağırlıklarının da öğrenilmesi sayesinde ilgili öznitelikler ilgisiz olanlardan ayrılabilmekte, bunun da ötesinde özniteliklerin bağıl önemleri de belirlenebilmektedir.

Bu projede öznitelik ağırlıklarının öğrenilmesinde genetic algoritmalardan yararlanılması da incelenmiştir. Bu amaçla, yeni bir çaprazlama operatorü de geliştirilmiştir.

Proje sırasında geliştirilen teknikler sayesinde çeşitli alanlarda çoklu kavram tanımlamalarının mekanik olarak öğrenilebilmesi ve her özniteliğin diğer özniteliklere göre önem derecesinin belirlenebilmesi mümkündür. Geliştirilen öğrenme tekniklerinin gerçek hayat problemlerine uygulanabilirliğinin sınanmasına yönelik olarak, biri EKG işaretlerinden aritmi karakteristiklerinin tanınması, diğeri de bir grup deri hastalığının histopatolojik tanımı için olmak üzere iki tıbbi verikumesi oluşturulmuştur. Proje kapsamında geliştirilen teknikler hem bu verikumeleri ve hemde UCI-Repository'deki standard verikumeleri üzerinde sınanmış ve çok başarılı sonuçlar alınmıştır.

Anahtar Kelimeler: Makina Öğrenmesi, Kavram Öğrenme, Genetik Algoritmalar, Sınıflandırma

9- Proje ile ilgili Yayın/Tebliğlerle ilgili Bilgiler

- Gülsen Demiröz, H. A. Güvenir, and Nilsel Ilter, Differential Diagnosis of Eryhemato-Squamous Diseases using Voting Feature Intervals in New Trends in Artificial Intelligence and Neural Networks(Proceedings of TAINN'97) Turhan Ciftcibasi, Mustafa Karaman and Volkan Atalay (Eds.), Kizilcahamam,Ankara, (May 22-23, 1997), 190-194.
- Gülsen Demiröz, and H. A. Güvenir, Classification by Voting Feature Intervals in Proceedings of 9th European Conference on Machine Learning Maarten van Someren and Gerhard Widmer (Eds.) Springer-Verlag, LNAI 1224, Prague, Czech Republic, (April 23-25, 1997), 85-92.
- G. Demiröz, and H. A. Güvenir, Genetic Algorithms to Learn Feature Weights for the Nearest Neighbor Algorithm in Proceedings of the 6th Belgian-Dutch Conference on Machine Learning (BENELEARN-96), H. J. van den Herik and T. Weijters (Eds.), Universiteit Maastricht, The Netherlands, (September 20, 1996), 117-126.
- A. Akkus, and H. A. Güvenir, K Nearest Neighbor Classification on Feature Projections, in Proceedings of ICML'96, Lorenza Saitta (Ed.), Morgan Kaufmann, Bari, Italy, (3-6, July, 1996), 12-19

10- Bilim Dalı:
Doçentlik B. Dalı Kodu: ISIC Kodu:
Uzmanlık Alanı Kodu:

11- Dağıtım (*): ❒ Sınırlı ■ Sınırsız
Tüm Üniversitelerin Bilgisayar Mühendisliği Bölümlerine ve Tıp Fakültelerine dağıtılabilir.

12- Raporun Gizlilik Durumu: ❒ Gizli ■ Gizli Değil