İris Çiçeğinin Sınıflandırılması
Classify Iris Flowers

Fatih KAPLAN
Elektrik Bilgisayar Mühendisliği
KTO Karatay Üniversitesi
Konya, Türkiye
fatih.kaplan@karatay.edu.tr

***Özetçe*— Bu çalışmada, iris çiçeğinin sınıflandırılması ile ilgili çalışmalar incelenmiş ve kullanılan algoritmalara göre sonuçlar sunulmuştur. Iris çiçeğinin ne olduğu ve neden makine öğrenmesi konularında sıklıkla kullanılageldiğinde bahsedilmiştir. Weka yazılımı ile gerçekleştirilen sınıflandırma işleminde "Cross-validation" test seçeneği kullanılmıştır. "Percentage Split" test seçeneği ile arasındaki farklara yer verilmiştir. "Percentage" yani yüzdelik dilime göre Weka'nın kendisinin test ve eğitim verilerini otomatik olarak ayarlayabildiği test türünün kullanımına değinilmiştir. Sınıflandırma algoritmalarından iki tanesi ile işlem gerçekleştirilmiş ve sonuçlar kaydedilmiştir. Aralarındaki farklar ve sonuçlardaki değişimlerden de bahsedilmiştir. İlk olarak sınıflandırma algoritmalarından "Lazy IBK" algoritması kullanılmıştır. Sınıflandırma algoritmasının yapılandırılabilen özelliklerinden komşuluk ilişkisi 3 iken sınıflandırmada 49 adet başarılı yerleştirme gözlemlenmiş ve hata 2 olarak başarı oranı 96% ve hata oranı 3,9216% iken komşuluk ilişkisi (KNN) 5 olduğu zaman "Confision Matrix" ekranında hata sayısının azaldığı gözlemlenmiştir.**

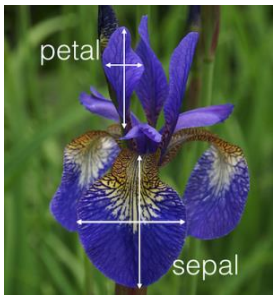
Anahtar Kelimeler — Sınıflandırma algoritması; makine öğrenmesi; komşuluk ilişkisi; Confision Matrix; Doğru sınıflandırılmış kayıtlar; Hatalı sınıflandırılmış kayıtlar;

***Abstract*— In this study, the classification of iris flowers was investigated. Results were recorded according to the algorithms used. It is mentioned that iris flower is what it is and why it is used frequently for learning machine. The cross validation test option has been used in the classification process performed with the weka software. The difference between the test method and the test method that divides it into percentages is given. In the percentile slice test option, weka software automatically sets the data as test and training data. Two of the classification algorithms found in the Weka software were performed and the results were recorded. The differences between them and the changes in the results. First, the Lazy IBK algorithm is used in the classification algorithms. 49 neighbors were found to be successful in the classification of the neighbors relation 3 from the configurable features of the algorithm. While in this way error 2, success 49 success rate 96% error rate is 3,9216%. When the neighborhood relation (KNN) is 5, the number of errors is observed to decrease.**

Keywords — Machine Learning; Classify Algorithms; Neighbor Relation; Correctly Classified Instances; Confision Matrix; Error Classified Instances.

I. GİRİŞ

Makine öğrenmesi konularında popüler olan veri setlerinden birisi de iris çiçeğine ait olan veri setidir. İris çiçeğine ait üç türden 50'şer tane toplamda 150 tane olmak üzere 1936 yılında bilim insanı Edgar ANDERSON tarafından toplanmış bir çiçektir. İris çiçeğinin Setosa, Versicolor ve Virginica türleri bulunmaktadır. Bilim insanı toplamış olduğu 150 çiçeğin üç türünün de üst ve alt çiçek yapraklarını ölçmüş ve bu ölçümden dört nitelikli ve 150 elemanlı bir veri seti elde etmiştir. Bu nitelikler ise sepal-length (alt-çanak yaprak uzunluğu), sepal width (alt-çanak yaprak genişliği), pedal-length (üst-taç yaprak uzunluğu) ve petal-width (üst-taç yaprak genişliği) şeklindedir.



Şekil 1 İris Çiçeği

Elde edilen bu veri seti makine öğrenmesi alıştırmalarında çok sıklıkla kullanılmaktadır. Makine öğrenmesi çalışmalarında olduğu gibi bu çalışmada da amaç iris çiçeğinin dört niteliğini kullanarak hangi türe ait olduğunu tahmin etmektir. Bir çeşit sınıflandırma problemidir.

Bu problemin çözümünde K-en yakın komşu yaklaşımı kullanılmıştır. Sınıfı bilinmeyen yeni bir çiçeğin yaprak ölçüleri kendisine en yakın K adet çiçekten en çok hangi sınıfa mensup ise o sınıftadır cevabını verebilecek yaklaşımdır.

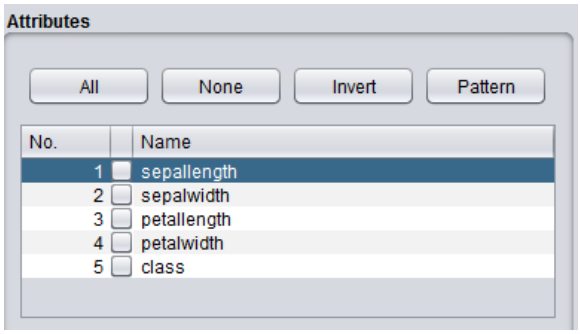
II. KONUYLA İLGİLİ YAPILMIŞ ÇALIŞMALAR

Konuyla alakalı yapılmış çalışmalardan Hindistan'da bir üniversite olan Amity School'dan Poojitha V, Shilpi Jain, Madhulika B hadauria ve Anchal Garg isimli öğrencilerin "A collocation of IRIS Flower using Machine Learning and Neural network" isimli makalelerinde iris çiçeğinin sınıflandırılmasında bir grup bilgiyi kümeler olarak bilinen benzer alt sınıfların bir grubuna bölme prosedürüdür şeklinde açıklamışlardır. Aynı çalışma içerisinde önceden tanımlanmış olan herhangi bir sınıfa sahip olmayan denetimsiz bir bölümü anlamak için farklı algoritmaların kullanılması gerektiğine değinilmiştir. Bir başka çalışma olan "Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001, June). Constrained k-means clustering with background knowledge. In ICML" makalelerinde ise daha kaliteli kümelerin çok iyi bir kümeleme yöntemi ile elde edilebileceğinden bahsedilmiş ve bununla ilgili çalışmalar yapmışlardır. Sınıf içi benzerliklerin yüksek ve sınıflar arasındaki benzerliklerin düşük olması gerektiğine değilnilmiştir.[7] Bir başka çalışmada ise iris çiçeğinin sınıflandırılmasında bazı alanların çok geniş uygulamalara sahip olduğundan bahsedilmiştir. Örneğin bir toprak algılama veritabanında karşılaştırabilir alanların tanınabilir kanıtları gibi. [6]

III. WEKA'DA SINIFLANDIRMA UYGULAMASI

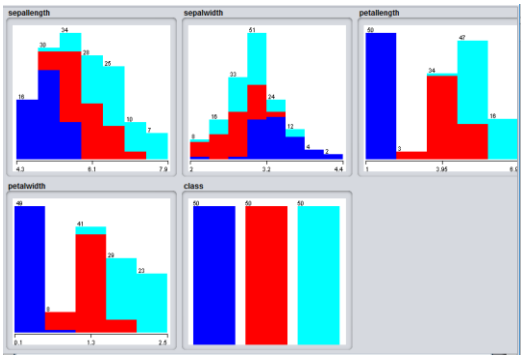
İris çiçeğinin sınıflandırılması uygulamasında toplam 150 adet veri (data sayısı) bulunmaktadır. Öznitelik olarak ise (attibute) 4 tane attibute bir tane de class attribute içerisinde yer almakta olup toplamda 5 tane özellik bulunmaktadır. Weight ise 150 tanedir. Seçilen dataset (iris data) içerisinde bulunan kayıtların ve özelliklerin sayısından bahsettik fakat bu data seti nasıl çalıştırıp sınıflandıracağız ve bunun için ne gerekiyor biraz da bundan bahsetmek gerekirse sınıflandırma işlemi için weka yazılımı kullanılacaktır. Weka yazılımı .csv formatındaki verileri de kendisi otomatik olarak algılayabiliyor olsa da varsayılan dosya formatı olan arff dosya formatı ile çalışmak daha efektif olacaktır. Arff olmayan dosyaları da arff dosyasına dönüştürmek mümkündür. Herhangi bir metin editörü ile veri seti açıldığı zaman veri setlerinin üstünde veri setine dair olmak üzere relation ve class eklemeleri yapılarak dosya arff dosyası biçimine getirilir. Uzantı olarak da .arff olarak kaydedildiğinde weka yazılımının algılayacağı bir veri seti haline gelmiş olur. İris data set sınıflandırma problemine yönelik bir veri setidir.

Öncelikle Weka programı açıldığında kullanıcıyı karşılayan diyalog kutusunda en üstte yer alan buton olan "Explorer" butonu ile weka programına giriş yapılır. Gelen pencerede "Preprocess" sekmesi açık olarak gelir ve burada yer alan open file butonu ile daha önce hazırlamış olduğumuz iris.arff isimli iris veri seti programa yüklenir. Veri setinin programa yüklenmesi ile birlikte Current Relation penceresinde veri setine ait bilgiler hemen görüntülenir hale gelir. Buradan relation iris, attiributes 5, instances 150 ve sum of weights 150 alanlarını ve değerlerini görüntüleriz. Attibutes penceresinden ise kaç özellikleri sıralanmış bir şekilde aşağıdaki şekilde de olduğu gibi görüntülemiş oluruz.



Şekil 2 Weka'da Iris Attributes

Attibutes penceresindeki özelliklerden herhangi birisi seçildiği zaman açık olan pencerede "selected attribute" bölümünde seçilen özelliğe ait bilgiler ve hemen altında ise grafiksel olarak seçilen özelliğe ait bilgiler görüntülenir. Class özelliği seçildiği zaman "selected attricute" bölümünde çiçek çeşitleri olan iris-setosa, iris-versicolor ve iris-virginica ile sayıları görüntülenir. Bu pencere ilgili veri setini yükledikten sonra o veri seti ile ilgili bilgileri sayısal ve grafiksel olarak göstermektedir. Weka programı bizim için attributes penceresindeki özelliklerin değerlerini hesaplayabiliyor ve visualize yaparak grafiksel olarak renklerle gösterimi sağlıyor. Bu kolonları standart sapma ve ortalama değer ile hesaplayabilmektedir.



Şekil 3 Veri Setinin Visualize Gösterimi

Attribute Selection menüsü ile bir özelliğin diğer özellik ile hareket ettiği gibi durumlarda özelliklerin elenmesi veya yeni özelliklerin çıkarılması gibi hususlarda kullanılan bir menüdür. Select Attributes menüsünden bir yöntem seçildi. Yöntem olarak "InfoGainAttributeEval" yöntemi ve search method olarak "Best First" methodu seçildi. Search Method çıktısı aşağıdaki gibi sonuçlandı.

TABLO I. SEARCH METHOD ÇIKTISI

Search Method	**Best First**
Start set	no attributes
Search direction	forward
Stale search after 5 node expansions	
Total numberof subsets evaluated	12
Merit of best subset found	0.887

TABLO II. SELECTED ATTRIBUTES ÇIKTISI

Search Method	Selected Attributes
Selected attributes	3,4 : 2
petallength	
petalwidth	

Daha sonra sınıflandırma işlemine başlamak için "Classify" menüsü altından algoritma seçeneklerinden "Lazy/IBk" algoritması kullanılmıştır. Weka programında 4 çeşit test türü vardır. Bunlar "use training set, supplied test set, cross-validation, percentage split" test türleridir. "Percentage Split" test türünde Weka otomatik olarak test ve eğitim verisini yüzdelik olarak veri setinden ayarlayabilmektedir. İlk olarak Cross Validation test türünde gerçekleştirilen sınıflandırma işleminde varsayılan olarak komşuluk yakınlığı 1 dir. Bu sınıflandırmaya göre elde edilen sonuçta 150 kayıt sayısından 143 adet kayıt doğru olarak sınıflandırılmıştır. Başarı yüzdesi ise 95,333% 'dir. Diğer sonuçlar ise aşağıdaki gibidir.

TABLO III. DİĞER SONUÇLAR

Search Method	Other
Root mean squared error	0.1747
Kappa statistic	0.93
Mean absolute error	f0.0399
Relative absolute error	8.9763 %
Root relative squared error	37.0695 %

TABLO IV. CONFUSION MATRIX

Confusion Matrix	KNN 1
a b c <--	classified as
50 0 0 | a =	Iris-setosa
0 47 3 | b =	Iris-versicolor
0 4 46 | c =	Iris-virginica

kNN değeri 1 yerine 5 olunca ise sonuçlar aşağıdaki gibi değişiklik göstermiştir.

TABLO V. CONFUSION MATRIX

Confusion Matrix	KNN 5
Correctly Classified Instances 143	95.3333 %
Kappa statistic	0.93
Mean absolute error	0.0413
Root mean squared error	0.1664
Relative absolute error	9.3 %
Root relative squared error	35.2926 %
Total Number of Instances	150

Algoritma yine aynı kaldığında fakat test türü Cross Validation yerine Percentage Split olarak ayarlandığında ise sonuçlar aşağıdaki gibi farklılıklar göstermektedir.

TABLO VI. PERCENTAGE SPLIT

Percentage split	66%
Correctly Classified Instances 49	96.0784 %
Kappa statistic	0.9408
Mean absolute error	0.0323
Root mean squared error	0.1395
Relative absolute error	7.2511 %
Root relative squared error	29.5071 %
Total Number of Instances	51

TABLO VII. CONFUSION MATRIX

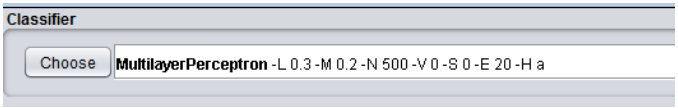
Confusion Matrix	KNN 1
a b c <--	classified as
15 0 0 | a =	Iris-setosa
0 19 0 | b =	Iris-versicolor
0 2 15 | c =	Iris-virginica

Doğru sınıflandırılmış veri sayısı 49 olarak hesaplanmıştır. Ve başarı oranı 96.078% olarak değişmiştir. Toplam kayıt 51 adet olduğundan doğru sınıflandırma 49 adet olduğu için hatalı sınıflandırma ise 2 adettir. Ve 3,9216% olarak hatalı sınıflandırma söz konusu olmuştur. Classify Choose textbox'ı üzerine tıklanınca açılan diyalog kutusundan ise kNN değeri ve birçok özelliği değiştirebiliriz. Yakın komşuluk değerini (kNN) 5 olarak güncelleyip tekrar sınıflandırma yaptığımızda ise sınıflandırmanın 5 yakın komşuya bakarak yapıldığından hataların sayısının azaldığı görülmektedir.

Karşılaştırma yapabilmek için yeni bir algoritma daha seçilerek sonuçlara ulaşılır. Ana ekrandan Choose butonuna tıklayarak açılan ekrandan Classifiers-functions kısmından bu sefer "MultilayerPerceptron" algoritması seçildi. Bu ekranda iken sınıflandırmaya başlamadan önce sınıflandırmanın Test Options sekmesinden modelin test yöntemi belirlendi. Test seçeneklerinden ilki "Use Training Set" eğitimde kullanılan verinin testte de kullanılması durumudur. İkinci test seçeneği olan "Supplied Test Set" türünde eğitim verisinden ayrı bir test verisi girilmesini beklemektedir. Üçüncü test türü "Cross Validate" türünde veri seti n parçaya ayrılarak her seferinde bir parça eğitimde, kalan n-1 parçanın tamamı ise test için kullanılarak sonuçta n test işleminin ortalaması alınmaktadır. Dördüncü test türünde "Percentage Split" veri setinin belirlenen yüzdesi test verisi olarak ayrılmakta ve testte

kullanılmaktadır. Sınıflandırmada test yöntemi olarak dördüncü test yöntemi yüzde değeri de 50% seçilerek start butonuna tıklandı. Bu işlem modelin oluşturulması, verinin eğitilmesi ve test edilmesi aşamalarını kapsamaktadır.

Classifier Output ekranı ikinci algoritmanın sınıflandırılmasında detaylıca incelendiğinde "Total number of instances" test için kullanılan örnek sayısını göstermektedir ve belirlenen şekilde 75 olarak görülmüştür. 75 test örneğinin gerçek sınıf ve tahmin edilen sınıf etiketlerinin dağılımı ise "Confusion Matrix" yani karışıklık matrisinde verilmiştir. Karışıklık matrisi 22 setosa çeşidinin tamamının doğru tahmin edildiğini, 26 versicolor çeşidinin yine tamamının doğru tahmin edildiğini, 27 virginica çeşidinin ise 3 tanesinin versicolor olarak yanlış tahmin edildiğini kalanların doğru tahmin edildiğini göstermektedir. Açık olan ekranda algoritmanın adının gösterildiği yerdeki modelin parametreleri yer almaktadır. Aşağıdaki şekilde de gösterilen bu parametreler aşağıdaki gibi açıklanmıştır.



Şekil 4 Algoritma Parametreleri

-L öğrenme oranını belirler ve 0-1 arasında değerler alır. -M geri yayılma algoritmasında momentum değerini belirler 0-1 arasında değerler alır. -H saklı katmanı gösterir. Node sayılarını belirler.

IV. KULLANILAN ALGORİTMALAR

A. *K-Means Algoritması*

K aracı, kesinlikle iyi bilinen gruplama sorunlarından birisini çözmek için kullanılan çok zor olmayan öğrenme prosedüründen bir tanesidir. Bu algoritma ile bölümlenmiş kümeleme yöntemi kullanılmaktadır. Sistem belirli bir grup kimliğini belirlemek için temel ve basit bir yaklaşımı geçerli kılar. K-aracı yalnızca sayısal bilgiler ile çalışmaktadır. Algoritmada rastgele bir sayı (K) küme merkezi seçilir. Her madde en yakın küme merkezine atanır. Her küme merkezini atanmış öğelerin ortalamasına taşınır. Yakınsama oluncaya kadar 2 ve 3.adımlar tekrarlanır.

B. *Perceptron Algoritması*

Algıyı modellemede kullanılmıştır. Duyular, eşleştirme ve cevap şeklinde üç parçadan oluşur. Öğrenim sadece A birimlerinden R birimlerine ağırlıklar üzerinde meydana gelir. Her bir **R** birimi n tane A biriminden giriş alır. Ağın öğrenebilmesi için eğitim seti adı verilen ve örneklerden oluşan bir sete ihtiyaç vardır. Çok katmanlı ağın öğrenme kuralı en küçük kareler yöntemine dayalı "Delta Öğrenme Kuralı"nın genelleştirilmiş halidir.

V. SONUÇLAR

Veri öğelerinin optimal kümelenmesi geleneksel yaklaşımlara göre sinir ağı kümeleme yoluyla başarılabilir, burada küme varyasyonları, veri kümelerinden merkezlerin rastgele seçilmesi yerine kümelerin sayısının belirlenmesi için düşünülmüş, iç kümeler vasıtasıyla düşünülmüş, iç kümeler vasıtasıyla ağırlık merkezleri optimal kümeler için hesaplanır ve uygunluk mutasyonuna dayalı uzaklık vasıtasıyla ölçülür. Iris kümelemesi kurumsal bilgi işleminde veri kümelemesi için geleneksel kümeleşmenin karşılaştığı birçok sorunu, kısıtlamaları ve eksiklikleri çözer. Doğrusal diskriminant analizi ile kümelenme için geniş veri kümeleri için uygun hale gelir. Bu makale, makine için üç farklı kategorideki Iris veri kümesi kapasitesini veren sınıflandırma ve kümeleme için Weka'da bulunan iki farklı algoritma ile mevcut iris çiçeği veri kümesini uygulamak için yapılan bir incelemedir. Kullanıcı makineye Iris'in hangi sınıfa uyduğunu söylemek zorunda değilse, makine hepsini hatırlayabilir.

KAYNAKLAR

[1] G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, Chipman, H., Hastie, T. J., & Tibshirani, R. (2003).

[2] Elena, M. (2013). Fuzzy C Means Clustering In Matlab. The 7th International Days of Statistics and Economics, Prague, 905-914. (introduction about clustering). Rudnicky, A. I., Polifroni, Thayer, E H., and Brennan, R. A. "Interactive problem solving with speech", *J. Acoust. Soc. Amer., Vol. 84, 1988, p S213(A).*

[3] Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery, 2(3), 283- 304

[4] Su, M. C., & Chou, C. H. (2001). A modified version of the K-means algorithm with a distance based on cluster symmetry. IEEE Transactions on Pattern Analysis & Machine Intelligence, (6), 674-680

[5] Kruskal, J. B. (1976). The relationship between multidimensional scaling and clustering. Classification and clustering, 17-44.

[6] Azzag, H., Venturini, G., Oliver, A., & Guinot, C. (2007). A hierarchical ant based clustering algorithm and its use in three real-world applications.European Journal of Operational Research, 179(3), 906- 922.

[7] Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001, June). Constrained k-means clustering with background knowledge. In ICML (Vol. 1, pp. 577-584).

[8] http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7508047

[9] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[10] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.