Dempster'in Birleştirme Kuralı ile Sınıflandırma Algoritmalarının Birleştirilmesi

H. Aygün
haygun@dho.edu.tr
Deniz Harp Okulu
Bilgisayar Mühendisliği Bölümü

E. Adalı
adali@cs.itu.edu.tr
İstanbul Teknik Üniversitesi
Bilgisayar Mühendisliği Bölümü

Özetçe

Bu çalışmada sınıflandırma sonuçlarını iyileştirmek için yeni bir yöntem önerilmektedir. Önerilen yöntem Dempster'in Birleştirme Algoritmasını kullanarak, farklı sınıflandırma algoritmalarından elde edilen sonuçların birleştirilmesidir. Dempster'in Birleştirme Algoritmasının kullanımıyla yapılan birleştirme işleminin, birleşimde kullanılan her bir sınıflandırma algoritmasından daha başarılı sonuçlar verdiği ortaya konmuştur. Ayrıca önerilen yöntemin mevcut birleşik algoritmalardan daha başarılı olduğu da gösterilmiştir. Sınıflandırmanın doğruluğunu artırmak amacıyla, birleştirme işleminde güven derecesi kullanımı önerilmiştir. Güven derecesi kullanımının daha doğru sınıflandırma gerçekleştiği gösterilmiştir.

Absract

In this study, we propose a method for combining classifiers in order to improve the performance of classification. The method consists of combining the classification results using Dempster's Rule of Combination, considering the classifier outputs as beliefs. In the combination we use some of the existing classification algorithms. We do experiments with different data sets to support our proposed method and we conclude that combining the classifier outputs using Dempster's Rule of Combination gives better classification results than each of the classification algorithms. We also use degree of confidence during the combination in order to improve the accuracy of the classification. Experimenting with several data sets shows that the employment of degree of confidence during combination results in more precise classification results.

1. Giriş

1982'e en iyi satan kitaplar arasında yer alan *Megatrends* kitabının yazarı John Naisbitt kitabında "Veri denizinde boğuluyoruz, ancak değerli bilgiye açız. Bu seviyedeki bir bilginin mevcut araçlarla yönetilmesi mümkün değildir. Kontrolsüz ve organize edilmemiş bilgi, bilgi toplumu için bir kaynak oluşturamaz, aksine bir düşman olur." ifadesini kullanmıştır. Bugün bu ifade daha fazla anlam kazanmıştır. Büyük miktardaki organize edilmemiş veri arasından değerli bilginin bulunup çıkarılmasına ihtiyaç duyulmaktadır. Bu sorunun iki çözümü vardır: veri madenciliği ve veri füzyonu.

Veri madenciliği önceden bilinmeyen fakat yararlı bilginin büyük miktardaki veri arasından bulunup çıkarılmasıdır. Veri madenciliği algoritmaları pazar analizi, risk analizi, kredi kartı sahtekarlıklarının belirlenmesi, metin ve web madenciliği gibi alanlarda yaygın olarak kullanılmaktadır. Örneğin pazar analizinde, veri madenciliği hedef pazar için aynı özelliklere sahip müşteri gruplarını bulmaya çalışır, müşterilerin zaman içerisindeki satın alma örüntülerini belirler ve ürün satışları arasındaki ilişkileri bulmaya çalışır.

Veri füzyonu ise farklı sensörlerden gelen bilgilerin birleştirilmesi işlemidir. Veri füzyonu algoritmaları, savunma sektöründe hedef tespiti, hedef kimlik tespiti amacıyla istihbarat, keşif ve gözetleme operasyonlarında kullanılmaktadır. [1]'de veri füzyonunun tanımı şu şekilde verilmiştir: İyileştirilmiş mevki ve kimlik tespiti yapmak, zamanında ve tam durum muhakemesi yapmak amacıyla bir veya birden fazla kaynaktan gelen bilginin ilişkilendirimesi ve birleştirilmesini içeren

bir prosestir. Sözkonusu proses, iyileştirilmiş sonuçlar elde etmek amacıyla tahmin ve değerlendirmelerinde sürekli olarak düzeltmeler yapar.

Veri madenciliği ve veri füzyonu birbirini tamamlayan prosesler olmasına rağmen, araştırmacılar bu iki alanda birbirinden bağımsız olarak, herhangi bir ilişkiye girmeden çalışmaktadırlar. Performansı artırmak için bu alanlarda kullanılan teknikleri birleştiren çok az sayıda çalışma mevcuttur. [2]'daki çalışma, görüntü verisi, yersel veri, video görüntüleri, istatistiksel veri takımları, anahtar kelimeler içeren veri içinde gizlenmiş hedeflerin tespiti ve sınıflandırması işlemlerinin iyileştirilmesi için veri madenciliği ve veri füzyonu tekniklerini birleştirilmesi yollarını araştırmaktadır.

Bu bildiride, daha iyi sınıflandırma sonucu elde etmek amacıyla, farklı sınıflandırma algoritmalarından elde edilen kanıtların Dempster'in Birleştirme Kuralı ile birleştirilmesinden oluşan bir yöntem önerilmektedir.

Bildirinin organizasyonu şu şekildedir: 2. Bölümde Dempster-Shafer Yöntemi, veri madenciliği sınıflandırma teknikleri ve önceki çalışma hakkında temel bilgi verilmektedir. 3. Bölümde önerilen birleştirilme yöntemi sunulmaktadır. UCI Veri Kütüphanesinden alınan farklı veri takımları üzerinde yapılan deneyler 4. Bölümde yer almaktadır.5. Bölümde de bildiri sonlandırılmaktadır.

2. Ön Bilgi

Bu bölümde Dempster-Shafer Yöntemi, sınıflandırma algoritmaları ve önceki çalışma hakkında özet bilgi verilmektedir.

2.1 Dempster-Shafer Yöntemi

Dempster-Shafer Yöntemi hipotezlere olasılık aralıkları atar. Dempster-Shafer Yönteminin girdileri farklı duyarga raporlarından elde edilen temel olasılık atama fonksiyonlarıdır (bpa). Duyargaların bpa fonksiyonlarını birleştirerek yeni bpa fonksiyonları elde etmek mümkündür. Bpa'lar birleştirildikten sonra hipotezler için olasılık alt ve üst sınırları ([Bel, Pla]) hesaplanır.

Matematiksel olarak ifade edecek olursak, m_1 ve m_2 iki bağımsız duyarga olsun. Θ gözlemlenen durumların kümesi olsun. Bilgi kaynaklarının verdiği bilgiler, Θ'nın kuvvet kümesi 2^{Θ} üzerinde tanımlıdır. 2^{Θ}'nın herbir elemanına [0,1] aralığında bir sayı karşı getirilsin. Öyleki, bu sayıların toplamı 1 olmalıdır. Bu işlemin matematiksel ifadesi aşağıdaki şekildedir [3]:

$$m : 2^{\Theta} \rightarrow [0,1] \tag{1}$$

$$m(\phi) = 0 \tag{2}$$

$$\sum_{A \subset \Theta} m(A) = 1 \tag{3}$$

m : bpa fonksiyonu veya olasılık kütle fonksiyonu

Dempster'in Birleştirme Kuralı:

$$m(A) = \frac{\sum_{B \cap C = A} m_1(B) \cdot m_2(C)}{1 - \sum_{B \cap C = \phi} m_1(B) \cdot m_2(C)} \tag{4}$$

m1 ve m2'nin sıfırdan büyük değerlerine "odak elemanları" denir. Kural şu şekilde yorumlanabir: İki bilgi kaynağının zıtlaştığı yani $B \cap C = \Phi$ olduğu durumlarda, bu bilgiyi destekleyen TOA değerlerinin çarpımı $B \cap C \neq \Phi$ olanlara dağıtılır.

Belief (Bel), Plausibility (Pla), $A \subset \Theta$ olmak üzere:

$$Bel(A) = \sum_{B \subset A} m(B) \tag{5}$$

$$Pla(A) = \sum_{A \cap B \neq \phi} m(B) \tag{6}$$

Bel ve Pla fonksiyonları arasındaki ilişki;

$$Bel(A) \leq Pla(A) \tag{7}$$

Dempster'in Birleştirme Kuralı temel olasılık atama fonksiyonlarını birleştirir ve birleştirilen bilgiyi temsil eden yeni bir olasılık atama fonksiyonu oluşturur.

$$\upsilon = |bel - pla| \tag{8}$$

Dempster'in Birleştirme Kuralı ile temel olasılık atama fonksiyonları birleştirildiğinde (4), Θ'nın tüm alt kümeleri için benzerlik vektörleri kullanılarak olasılık kütle fonksiyonları hesaplanır. Daha sonra her bir alt küme için alt ve üst olasılık değerleri hesaplanır. Karar verme işlemi için her bir durumun bel değerleri birbiri ile karşılaştırılır. En büyük bel değerine sahip durum karar olarak seçilir. Değerler aynı veya birbirine çok yakınsa hesaplanan pla değerleri karşılaştırılır ve en büyük pla değerine sahip durum karar olarak seçilir.

Bu noktada, Dempster'in Birleştirme Kuralının kullanılabilmesi için duyargaların birbirinden bağımsız olması gerektiği gerçeğini de vurgulamalıyız.

2.2 Sınıflandırma Algoritmaları ve Önceki Çalışma

Sınıflandırma işlemi, önceden sınıflandırılmış örneklerin kullanılmasıyla bir model oluşturularak bir veri seti üzerinde sınıflandırma yapılmasını sağlar. Regresyon, Karar Ağacı, Bayes Sınıflandırması, En Yakın Komşu, Yapay Sinir Ağları gibi pek çok sınıflandırma yaklaşımları mevcuttur. Sınıflandırma yönteminin kullanıldığı alanlar müşteri ayrımı, kredi analizi, finansal marketlerde trend sınıflandırması ve iş modellemesi, tıbbi teşhis ve tedavi gibi alanlardır.

Sınıflandırma algoritmalarının birbirlerine karşı üstünlükleri ve eksiklikleri vardır. Farklı veri takımları üzerinde farklı performans gösterirler. Bazen bir örneği yanlış sınıflandırabilirler veya hiç sınıflandıramayabilirler.

Sınıflandırmanın başarısını artırmak için farklı sınıflandırıcıların sonuçları birleştirilebilir. [4]'te çevre bilgi sistemi otomasyonu amacıyla veri füzyonu ve veri madenciliği teknikleri birleştirilmektedir. Sınıf seviyesi füzyon/madencilikten elde edilen bilgi karar ağacı sınıflandırma işleminin girdisi olarak kullanılmaktadır.

[5]'te Dempster-Shafer teorisi tıp verisinin birleştirilmesi amacıyla kullanılmıştır. K-Komşu (kNN), Naive Bayes ve Karar Ağacı algoritmaları Dempser'in Birleştime algoritması ile birleştirilmiştir. Bu çalışmada sınıflandırıcı sonuçlarının birleştirilmesinden elde edilen iyileştirme 0.1% ile 1.1% arasındadır.

[6]'de "Ağırlıklı Dempster-Shafer kanıt birleştirme kuralı" kullanılmıştır. Onların yaklaşımında birleştirme sırasında duyargaların ağırlıkları kullanılmıştır.

[7]'de Dempster-Shafer çatısı altında çelişen bilgilerin ağırlıklı birleştirilmesi incelenmiş ve değiştirilen kuralın bazı özellikleri verilmiştir.

3. Önerilen Yöntem

Bu çalışmada, sınıflandırmanın performansını artırmak amacıyla farklı sınıflandırıcıların sonuçları Dempster'in Birleştirme Kuralı ile birleştirilmektedir. Güven derecesi veya güven faktörü sınıflandırma algoritmasının geçmişte gösterdiği başarı oranıdır. Sınıflandırıcıların güven faktörünü dikkate alarak sınıflandırıcı sonuçlarını birleştirmek mümkündür. Matematiksel olarak ifade etmek gerekirse; $\mathcal{L}'$ nin benzerlik vektörü olduğunu, C 'nin güven faktörü olduğunu kabul edersek yeni benzerlik vektörü $\mathcal{L}_{yeni}$ şu şekilde olur:

$$\mathcal{L}_{yeni}[a_{1,j}] = \mathcal{L}[a_{1,j}] * C$$

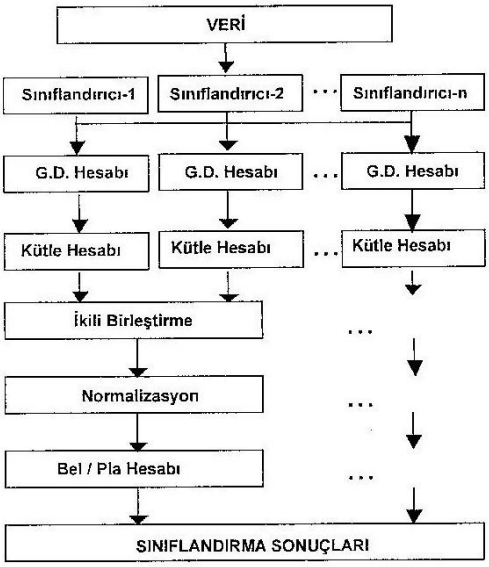
Önerilen yöntemde, önce farklı sınıflandırma algoritmaları ile sınıflandırma yapılır. Daha sonra her bir sınıflandırıcının güven derecesi hesaplanır. Ardından her bir sınıflandırıcı için olasılık kütle

fonksiyonları hesaplanır. Olasılık kütle değerleri çifter çifter Demsper'in Birleştirme Kuralı ile birleştirilir. Sonunda yeni *bel* ve *pla* değerleri hesaplanır. Önerilen yöntemin akış diyagramı Şekil-1'de verilmiştir.

Sınıflandırıcı sonuçlarından elde edilen kanıtlar kullanılarak, sınıflandırma algoritmalarının sonuçları Dempster'in Birleştirme Kuralı ile birleştirilmektedir. Önerilen yöntem WEKA'da mevcut basit ve bileşik sınıflandırma algoritmalarından daha iyi sonuçlar vermektedir.

4. Deneyler

Deneylerde Kaliforniya Irvine Üniversitesi Veri Kütüphanesinden (UCI Machine Learning Repository) [8][9]'den alınan 10 veri takımı kullanılarak sınıflandırma işlemi yapılmıştır. Deneylerde kullanılan veri takımları Tablo 1'de yer almaktadır.



Şekil 1. Önerilen Yöntem

Deneylerde WEKA sınıflandırma metotları [8] varsayılan değerleri ile kullanılmıştır.

Tablo 1. Deneylerde Kullanılan Veri

Veri	Kayıt Ad.	Nitelik
Autos	205	26
Breast-Cancer-Wisconsin	699	11
Heart-Disease-Cleveland	303	14
Heart-Disease-Hungary	294	14
Hepatitis	155	20
Iris	150	5
Labor	57	17
Soybean	683	36
Thyroid	215	6
Wine	178	14

WEKA sınıflandırıcılarının kullandığımız veri takımları üzerindeki başarısı Tablo 2'de yer almaktadır. Bu çalışmada dört temel sınıflandırma algoritması seçilmiştir: NaiveBayes, IB1, j48 ve OneR. Tablo 2'de görüldüğü gibi Naive Bayes veri takımları üzerinde en fazla ortalama başarıya sahiptir ve en büyük başarıyı Breast-Cancer-Wisconsin verisi üzerinde göstermektedir. Bu arada, IB1 en iyi başarıyı tiroid verisi üzerinde, j48 iris verisi üzerinde ve OneR da aynı şekilde iris verisi üzerinde göstermektedir.

Tablo 2. WEKA sınıflandırıcılarının UCI veri takımlarındaki başarısı (%)

Veri	Naïve Bayes	IB1	J48	OneR
Autos	79.51	94.15	89.75	83.41
Breast-Cancer-Wisconsin	97.42	95	94.85	92.70
Heart-Disease-Cleveland	56.43	54.45	52.47	52.47
Heart-Disease-Hungary	85.35	59.18	78.57	78.91
Hepatitis	70.32	66.45	58.70	61.93
Iris	96	95.33	96	94
Labor	89.47	82.45	73.68	75.43
Soybean	92.97	89.89	91.50	39.97
Thyroid	96.74	97.20	92.09	91.16
Wine	96.62	94.94	93.82	76.40
Ortalama	86.08	82.90	82.14	74.63

WEKA bileşik sınıflandırma algoritmalarından bazılarının UCI veri takımları üzerindeki ortalama başarısı Tablo 3'te verilmiştir. Bu bileşik algoritmaların seçilmesinin nedeni, birden çok algoritmayı birleştirmeleri nedeniyle önerilen yöntemle birebir karşılaştırma imkanını sunmalarından dolayıdır.

Tablo 3. WEKA Bileşik Sınıflandırma Algoritmalarının UCI Veri Takımları Üzerindeki Ortalama Başarısı

	Naïve Bayes + IB1	Naïve Bayes + J48	Naïve Bayes + OneR	IB1 + J48	IB1 + OneR	J48 + OneR
Grading	89.73	91.39	90.17	88.38	87.58	89.05
Multischeme	89.86	91.04	90.02	89.57	89.57	87.91
Stacking	90.33	91.61	89.78	83.14	78.38	87.66
Vote	89.54	91.39	77.90	89.60	84.96	81.86

Önerilen yöntemde, dört sınıflandırma algoritması Dempster'in Birleştirme Algoritması kullanılarak altı

farklı kombinasyonda birleştirilmektedir. Önerilen yöntemden elde edilen sonuçlar ile WEKA'da mevcut bazı bileşik algoritmaların sonuçlarının karşılaştırılması Tablo 4'te yer almaktadır.

Tablo 4. Önerilen Yöntem ile WEKA Bileşik Sınıflandırıcılarının UCI Veri Takımları Üzerindeki Başarısının Karşılaştırılması

	Naive Bayes + IB1	Naive Bayes + J48	Naive Bayes + OneR	IB1 + J48	IB1 + OneR	J48 + OneR
Dempster-Shafer	95.73	94.13	92.30	93.26	92.44	89.21
Grading	89.73	91.39	90.17	88.38	87.58	89.05
Multischeme	89.86	91.04	90.02	89.57	88.57	87.91
Stacking	90.33	91.61	89.78	83.14	78.38	87.66
Vote	89.54	91.39	77.90	89.62	84.86	87.85

Tablo 4'te görüldüğü gibi, sınıflandırma algoritmalarından elde edilen kanıtların Dempster'in Birleştirme Kuralı ile birleştirilmesinden oluşan önerilen yöntem, UCI veri takımları üzerinde, ortalama olarak mevcut WEKA bileşik algoritmalarından daha iyi bir performans sergilemektedir. Tablo 4'te görülmemekle birlikte, mevcut bileşik algoritmaların bir kısmı, bazı UCI veri takımları üzerinde önerilen yöntemden daha başarılıdır.

4. Sonuç

Bu çalışmada, sınıflandırmanın performansını artırmak amacıyla sınıflandırıcıların birleştirilmesi için bir yöntem önerilmektedir. Birleştirme işleminde Dempster'in Birleştirme Kuralı kullanılmaktadır. Önerilen yöntemde, aynı zamanda sınıflandırıcıların güven derecesi de dikkate alınmaktadır.

Dört farklı sınıflandırma algoritması altı farklı biçimde birleştirilerek UCI Veri Kütüphanesinden alınan 10 farklı veri seti üzerinde deneyler yapılmıştır. Deneylerin sonucu, önerilen yöntemin, mevcut basit ve bileşik sınıflandırıcılardan %5-15 oranında daha başarılı olduğunu göstermektedir. Bu başarının sebebi, Dempster'in Birleştirme Kuralında mevcut belirsizlik yönetimi yeteneğidir.

Kaynakça

[1] **White, Jr., F.E.**, Data Fusion Lexicon, Joint Directors of Laboratories, Technical Panel for C3, Data Fusion Sub-Panel, Naval Ocean Systems Center, San Diego, 1987.

[2] **E.Waltz** "Information Understanding: Integrating Data Fusion and Data Mining Processes", in workshop along with IEEE 1999 International Symposium on Signals and Systems.

[3] **G. Shafer**, A Mathematical Theory of Evidence, Princeton University Press, 1976.

[4] **M.Wachowicz and L.M.T.Carvalho** "Data Fusion and Mining for the Automation of a Space-Time Reasoning Process"

[5] **Mahajani, G. A., Aslandogan Y. A.**, "Evidence Combination in Medical Data Mining", Technical Report CSE-2003-23, Department of Computer Science and Engineering, University of Texas at Arlington, July 2003.

[6] **Wu H., Siegel2 M., Stiefelhagen R.,Yang J.**, "Sensor Fusion Using Dempster-Shafer Theory", IEEE Instrumentation and Measurement Technology Conference, Anchorage, AK, USA, 2002.

[7] **Basak J., Goyal Z., Kothari R.**, "A Modified Dempster's Rule Of Combination for Weighted Sources of Evidence", IBM Research Report, IBM Research Division, IBM India Research Lab, 2004.

[8] "Data Mining: Practical machine learning tools with Java implementations," by Ian H. Witten and Eibe Frank. Morgan Kaufmann, San Francisco, 2000.

[9] http://www.ics.uci.edu/~mlearn/ MLReposit.html