Naive Bayes Yönteminde Olasılık Çarpanlarının Etkileri
The Effects of Probability Factors in Naive Bayes Method

Umut Orhan[1], Kemal Adem[2]

[1]Elektrik-Elektronik Mühendisliği Bölümü
Gaziosmanpaşa Üniversitesi
umut.orhan@gop.edu.tr

[2]Mekatronik Mühendisliği Bölümü
Gaziosmanpaşa Üniversitesi
kemal.adem@gop.edu.tr

Özet

Bu çalışmada, basit yapısı ve yüksek başarısıyla bilinen Naive Bayes (NB) yönteminde kullanılan olasılık çarpanlarının sınıflandırmaya etkisi araştırılmıştır. Çalışmanın başarısı UCI veritabanından alınan iki sınıflı ve nümerik veri kümeleriyle test edilmiştir. Elde edilen sonuçlara göre sınıf olasılığı çarpanının sınıflandırmaya çoğu zaman yarar sağlasa da bazen zarar da verebildiği gösterilmiştir. Ayrıca NB yönteminin sınıflandırmayı olumlu-olumsuz etkileyen niteliklerin seçilmesinde kullanılabileceği üzerine bir öneri sunulmuştur. Sonuç olarak çalışmada, NB yöntemi hangi amaçla kullanılırsa kullanılsın genel başarı açısından veri bileşenlerinin etkilerinin NB yönteminde kullanılan her bir çarpanın sonuca katkısı aracılığıyla incelenmesi gerektiği vurgulanmıştır.

Abstract

In this study, the effect on classification of probability factors used in Naive Bayes method known its simple structure and high success is researched. The success of the study is tested by two-class and numeric datasets from UCI database. According to obtained results, it is shown that although the class probability factor is usually provided benefits, it can sometimes damage to the classification success. In addition, a suggestion is presented on the using of Naïve Bayes method in selection of the attributes which affect the classification success positively or negatively.

1. Giriş

Naive Bayes (NB) basit ve uygulama alanında üst seviyede doğruluk gösteren etkili bir istatistiksel tahminleme algoritmasıdır [1–2]. Temel olarak verideki niteliklerin sadece belirli bir sınıfa bağımlılığını koruyarak en uygun sınıflandırmayı gerçekleştirir. En çok metin sınıflandırma ve spam ayıklamada başarıları bilinen NB algoritmasını farklı konular üzerinde sınıflandırma yöntemi olarak kullanan birçok çalışma yapılmıştır [3–7]. Bu çalışmaların birinde Türkçe yazılı olan metnin sınıflandırılmasının en başarılı şekilde NB yöntemi kullanılarak yapıldığı gösterilmiştir [3]. Çalışmada Türkçe metinlerin sınıflandırılması için geliştirilmiş bir özellik çıkarma yöntemi anlatılmıştır. Metin sınıflandırmayla ilgili farklı bir çalışmada yazarı bilinmeyen bir dokümanın önceden belirlenmiş farklı yazarlar içerisinden hangisine ait olabileceği üzerine deneyler yapılmış ve deney sonuçlarına göre NB yönteminin başarılı sonuçlara ulaştığı belirlenmiştir [4]. Bu sonuçlar NB yönteminin iyi bir metin sınıflandırma yöntemi olduğunu göstermektedir. Bir başka çalışmada yazılım projelerinde risklerin girdi olarak kullanıldığı bir eğitim modeli oluşturup bu modeldeki risk verileriyle modeli test etmek için Bayes yöntemi ile birlikte farklı sınıflandırma yöntemleri kullanılmıştır [5]. Bu yöntemler karşılaştırıldığında Bayes yöntemi, Yapay Sinir Ağı kullanan yöntem kadar başarılı olamamıştır. Farklı bir çalışmada ise spam e-postaları tespit etme sürecinde NB yönteminin başarılı olduğu gözlemlenmiştir [6]. Finans sektöründe dolandırıcılık tespiti için NB yönteminden faydalanılarak çalışmalar yapılmıştır [7].

Bu çalışmada, NB tekniğinde kullanılan olasılık çarpanlarının sınıflandırmayı nasıl etkileyeceği üzerine çalışılmıştır. Çalışmanın başarısını test etmek için UCI veritabanından alınan yedi veri kümesi kullanılmıştır. Genel olarak çalışmanın geri kalanı şu şekilde düzenlenmiştir. İkinci bölümde çalışmada kullanılan veri kümelerinden, Bayes teoremi ve Naive Bayes sınıflayıcıdan bahsedilmiştir. Üçüncü bölümde yapılan çalışmanın veri kümeleri üzerindeki sınıflandırma deneyleri ve elde edilen başarı sonuçları verilmiştir. Son olarak dördüncü bölümde genel değerlendirmeler yapılarak çarpanların etkinliği üzerine eleştiri sunulmuştur.

2. Materyal ve Metot

2.1. Kullanılan Veri Kümeleri

Çalışmada yapılan deneyler için California Üniversitesi web sitesinde bulunan makine öğrenmesi veritabanından alınan iki sınıflı ve nümerik değerli yedi adet veri kümesi (Haberman, Ionosphere, Parkinsons, PimaDiabet, SpectHeart, Transfusion ve Wisconsin) kullanılmıştır. Bu veri kümelerinden Haberman, göğüs kanseri nedeniyle ameliyat geçirmiş

hastaların hayatta kalmaları üzerine yapılan çalışma değerlerini; Ionosphere, dünyanın iyonosfer tabakasından radarla alınan değerleri; Parkinsons, parkinson hastalığının tespit edilebilmesi için gerekli değerleri; PimaDiabet, kişilerin belirli özelliklerine bakılarak diabet hastalığının olup olmadığına dair bilgiler; SpectHeart, Single Proton Emission Computed Tomography (SPECT) cihazı ile kalp görüntülerinin normal veya anormal olduğuna dair bilgileri; Transfusion, kişilerin belirli değerlere göre kan nakli yapma durumuyla ilgili bilgileri; Wisconsin, göğüs kanseri teşhisinde kullanılan veri değerlerini içermektedir. Bu veri kümelerinin örnek ve nitelik sayıları Tablo1'de verilmiştir.

Tablo1. Çalışmada kullanılan veri kümelerinin örnek ve nitelik sayıları.

Dataset	Örnek	Nitelik
Haberman	306	3
Ionosphere	351	34
Parkinsons	195	22
PimaDiabet	768	8
SpectHeart	267	44
Transfusion	748	4
Wisconsin	683	9

2.2. Bayes Teoremi

Birbirinden bağımsız ve rasgele iki olayın (X ve Y) birbiri ardı sıra gerçekleştiği durumlarda bu iki olaydan birinin gerçekleşmesi durumunda ikinci olayın gerçekleşme olasılığı $P(X \cap Y)$ ifadesi ile gösterilebilir. Değişme özelliği sayesinde aşağıdaki çarpım kuralı iki farklı ifade ile yazılabilir.

$$P(X \cap Y) = P(X \mid Y)P(Y) = P(Y \mid X)P(X) \tag{1}$$

Bayes teoremi, rassal bir sürece bağlı olarak ortaya çıkan rasgele bir X olayı ile diğer bir rasgele Y olayı için koşullu olasılıklar ve marjinal olasılıklar arasındaki ilişkiyi tanımlar. Bu ilişkiyi ilk kez Thomas Bayes ortaya atmış ve aşağıdaki eşitliği önermiştir [8].

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} \tag{2}$$

Herhangi bir problemde gerçekleşmesi muhtemel bağımlı durumların olasılıkları yukarıda verilen Bayes eşitliği ile hesaplanır. Bu eşitlikte $P(X)$ ifadesi problemin girdi olasılığını, $P(Y)$ ifadesi olası çıkış durumunun olasılığını ve $P(Y \mid X)$ ifadesi ise daha önce gerçekleşen X girişine karşın Y çıkışı durumlarının olasılığını temsil etmektedir.

2.3. Naive Bayes Sınıflayıcı

Bayes Teoremi sınıflandırma amaçlı kullanılırken olası çıkış durumları içerisinde en büyük olasılığa sahip olan durum hedef sınıf olarak seçilir. Bunun matematiksel ifadesi aşağıdaki gibi gösterilir.

$$Y' = \arg\max_{y_j \in Y} P(Y = y_j \mid X) \tag{3}$$

Burada Y' değişkeni hedef sınıfı, y_j ifadesi olası j. çıkış durumunu ve X değişkeni de sınıfı belirlenecek olan giriş örneğini temsil etmektedir. Giriş örneğinin birden çok niteliğe sahip olduğu durumda Bayes formülü farklı bir forma dönüşür. Birçok özelliğin kesişimi görünümündeki veri örneği için hedef sınıf tahmininde tüm nitelikler için koşullu olasılıkların çarpımı hesaplanmalıdır.

$$P(x_1, x_2, ..., x_m \mid y_j) = \prod_{i=1}^{m} P(x_i \mid y_j) \tag{4}$$

Burada X giriş örneği, m adet nitelikten oluşmakta ve $X=(x_1,x_2,\ldots,x_m)$ şeklinde ifade edilmektedir. Naive Bayes (NB) sınıflayıcı ile Bayes Teoremi hesaplarında dikkat edilmesi gereken en önemli fark, sınıflayıcının olasılık değerinden ziyade hedef sınıfı bulmaya odaklanmasıdır. Bayes Teoreminde paydada bulunan değer, tüm hedef durumların olasılık hesaplarında ortak olduğundan NB sınıflayıcıda ihmal edilir. Bu durumda çok niteliğe sahip girişlerde NB formülü aşağıdaki şekle dönüşür.

$$Y' = \arg\max_{y_j \in Y} \left(P(Y = y_j) \prod_{i=1}^{m} P(X = x_i \mid Y = y_j) \right) \tag{5}$$

3. Nümerik Analiz

Bu bölümde, önceki bölümde verilen yöntem ve veri kümeleri kullanılarak yapılan deneylerin sonuçları değerlendirilmiştir. Sınıflandırma deneylerinin başarı değerlendirmesinde en sık kullanılan ölçüt Toplam Doğru Sınıflandırma (TDS) oranıdır. TDS oranı, sınıflayıcının doğru sınıflandırdığı örnek sayısının verideki toplam örnek sayısına bölünmesiyle hesaplanır.

NB yönteminin temelini oluşturan Bayes kuralından gelen sınıf olasılıkları çarpanı, veri kümesinin hangi sınıftan kaç örnek içerdiğiyle ilişkilidir. Bu çarpan, veri kümesini hazırlayan kişinin tercihine göre istatistiksel sınıflandırmayı etkileyebilmektedir. Bu çarpanın NB yöntemi üzerine etkisi kullanılan yedi veri kümesi üzerinde incelenmiş ve sonuçlar Tablo2'de gösterilmiştir.

Tablo2. Sınıf olasılığı çarpanının NB yönteminde sınıflandırma başarısına etkisi.

Dataset	TDS	TDS*
Haberman	74.51	76.14
Ionosphere	82.91	82.91
Parkinsons	69.74	69.74
PimaDiabet	76.17	75.13
SpectHeart	69.29	67.42
Transfusion	75.13	72.99
Wisconsin	96.63	96.49

(*: Sınıf olasılığından bağımsız TDS oranı)

Tablo2 incelendiği zaman sınıf olasılığı çarpanının sınıflandırma başarısını sadece Haberman veri kümesinde olumsuz etkilediği söylenebilir. Ionosphere ve Parkinsons veri kümelerinde bu çarpanın belirgin bir olumlu veya olumsuz etkisinden bahsedilemez. Diğer dört veri kümesinde (PimaDiabet, SpectHeart, Transfusion ve Wisconsin) ise sınıf olasılığı çarpanı varlığının sınıflandırma başarısını olumlu yönde etkilediği gözlenmiştir.

Sınıf olasılığının sınıflandırmadaki başarı oranını etkilediği gibi verideki her bir niteliğin olasılık çarpanının da başarı

oranını nasıl etkilediğini inceleyebiliriz. Böyle bir analiz ile aynı zamanda örüntü tanıma problemlerinde özellik seçme aşaması için kullanılabilecek yeni bir yaklaşım sunulabilir. Bunun için NB yöntemiyle veri kümeleri tekrar sınıflandırılmış ve her sınıflandırmada sırasıyla her bir nitelik hariç tutulmuştur. Dolayısıyla her veri kümesi, içerdiği nitelik sayısı ile ilişkili olarak tekrar sınıflandırılmış ve elde edilen en düşük başarı oranları (Min TDS), en yüksek başarı oranları (Max TDS) ve tüm nitelikler kullanılarak elde edilen başarı oranları (TDS) Tablo3'te gösterilmiştir.

Tablo3. Özellik seçme için elde edilen başarı oranları.

Dataset	Min TDS	Max TDS	TDS
Haberman	73.53	74.84	74.51
Ionosphere	81.20	84.61	82.91
Parkinsons	69.23	70.77	69.74
PimaDiabet	70.83	76.82	76.17
SpectHeart	66.29	70.79	69.29
Transfusion	74.46	76.60	75.13
Wisconsin	95.90	96.63	96.63

Tablo3 incelendiğinde Ionosphere veri kümesinde niteliklerden birinin hariç tutulmasıyla başarı %82.91'den %81.20'ye düşmüş ve böylece bu niteliğin varlığının sınıflandırmaya olumlu katkısı olduğu sonucu çıkmıştır. Yine aynı veri kümesinde farklı bir niteliğin çıkarılmasıyla başarı oranı %84.61'e yükselmiş ve bu niteliğin de sınıflandırmaya olumsuz etkisi olduğu belirlenmiştir. Wisconsin veri kümesini incelediğimizde bir niteliğin çıkarılmasıyla başarı oranı %96.63'den %95.90'a düşmüş ve yokluğunun sınıflandırmaya zarar vermesi dolayısıyla bu niteliğin kullanılması gerektiği sonucu çıkmıştır. Yine aynı veri kümesinde herhangi bir niteliğin çıkarılmasıyla başarı oranı artırılamamış ve bu niteliklerin bireysel olarak varlığının sınıflandırmaya ne yarar ne de zarar verdiği yorumlanmıştır. Yapılan bu analiz ile her niteliğin sınıflandırma başarısına etkisi araştırılmıştır. Bu özelliğiyle NB tekniğinin bir özellik seçme yöntemi gibi kullanılmasının yararlı olabileceği düşünülmektedir.

4. Sonuçlar

Bu çalışmada, Naive Bayes (NB) tekniğinde kullanılan olasılık çarpanlarının sınıflandırma başarısına etkisi incelenmiştir. Çalışmada yapılan deneyler için makine öğrenmesi yöntemlerinde sıklıkla kullanılan yedi farklı veri kümesi tercih edilmiştir. Deneylerde, Bayes kuralından gelen sınıf olasılığı çarpanının veri kümelerinin çoğunda sınıflandırma başarısını olumlu yönde etkilediği gösterilmiştir. Ayrıca veri kümelerindeki her bir niteliğin NB ile sınıflandırmaya katkısı incelenmiş ve bazı niteliklerin sınıflandırma başarısına olumlu bazılarının olumsuz etkisi yanında bazı niteliklerin de başarıya herhangi bir etkisi olmadığı gözlenmiştir. Sonuç olarak, NB yöntemi ister sınıflandırma aracı olarak ister örüntü tanıma problemlerinde özellik seçme yöntemi olarak kullanılsın, veri içerisindeki özelliklerin etkinlikleri NB yönteminde kullanılan çarpanların sistemin genel başarısına katkısının araştırılması ile tespit edilmelidir.

5. Kaynaklar

[1] Lewis, D. D., "Naive Bayes at forty: The independence assumption in information re-trieval", *In Proceedings of the Tenth European Conference on Machine Learning*, 1998, 4-15.

[2] Hand, D. J. ve Yu, K., "Idiot's bayes: Not so stupid after all?", *International Statistical Review*, 69, 3, 385-398, 2001.

[3] Yıldız, H. K., Gençtav, M., Usta, N., Diri, B., Amasyalı, M. F., "Metin Sınıflandırmada Yeni Özellik Çıkarımı", *15. Sinyal İşleme ve İletişim Uygulamaları Kurultayı*, 2007.

[4] Amasyalı, M. F., Diri, B., Türkoğlu, F., "Farklı Özellik Vektörleri ile Türkçe Dokümanların Yazarlarının Belirlenmesi", *Fifteenth Turkish Symposium on Artificial Intelligence and Neural Networks*, 2006.

[5] Chulani, S., Boehm, B., ve Steece, B., "Bayesian Analysis of Empirical Software Engineering Cost Models", IEEE Transactions on Software Engineering, 24, 4, 573-583, 1999.

[6] Eryiğit, G., Tantuğ, C., Adalı, E., "Yaramaz E-Postaların Süzülmesinde Karar Destek Makineleri, Naive Bayes ve Bellek Tabanlı Öğrenme Yöntemlerinin Karşılaştırılması", *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 1, 27-36, 2005.

[7] Özbay, E., "Finans Sektöründe Veri Madenciliği ile Dolandırıcılık Tespiti", *Selçuk Üniversitesi Yüksek Lisans Tezi*, 2007.

[8] Bishop, C. M., "Pattern Recognition and Machine Learning", *Springer*, 2006.