Makine Öğrenmesi Güz 2017 2. Ödevi

“Weka’da Yapay Nöron Ağları ve Destek Vektör Makineleri ile El Yazısı Tanıma”

14505013 İhsan Ozan YILDIRIM

Yıldız Teknik Üniversitesi

Not

Bu çalışmanın dosyalarına aşağıdaki adresten erişilebilir.

https://github.com/ozantoteles/wekaannsvm

Özet

Bu ödevde Yapay Nöron Ağları ve Destek Vektör Makineleri yöntemlerinin el yazısı sınıflandırmadaki başarısı değerlendirilmiştir. Veri seti olarak http://yann.lecun.com/exdb/mnist/ adresinden edinilmiş el yazısı karakterleri ve 0-9 arası karşılıklarını gösteren piksel değerleri tablosu kullanılmıştır. Ödev sonucunda her iki yöntemin başarıları benzer seviyelerde çıkmış ancak Destek Vektör Makineleri yöntemi zaman performansı ile öne çıkmıştır.

Anahtar Kelimeler: Makine Öğrenmesi, Yapay Nöron Ağları, Görüntü Sınıflandırma, El Yazısı Tanıma, WEKA, Destek Vektör Makineleri

14505013 İhsan Ozan YILDIRIM

Weka'da Yapay Nöron Ağları ve Destek Vektör Makineleri ile El Yazısı Tanıma

Giriş

WEKA makine öğrenmesi amacıyla Waikato Üniversitesi'nde geliştirilmiş ve "Waikato Environment for Knowledge Analysis" kelimelerinin baş harflerinden oluşmuş yazılımın ismidir[1]. Birçok başka makine öğrenmesi algoritması için hazır kütüphaneler ve arayüz sunmasıyla birlikte bu ödevde ilgilenilen Yapay Nöron Ağları ve Destek Vektör Makineleri için de kullanılabilecek bir araçtır. Her iki yöntem de eğiticili (supervised) öğrenme yöntemleridir.

Yapay Nöron Ağları insan vücudunda bilginin işlenmesine ve aktarılmasına yarayan sinir hücrelerini taklit etmek fikrinden ortaya çıkmış bir makine öğrenmesi yöntemidir. Yapay Nöron Ağları'nda öğrenme biyolojik nöronlardakine benzer şekilde her yeni iterasyonda birbirine bağlı nöronlara aktarılacak verilerin ağırlıklarının sonuç yakınsayana kadar değiştirilmesiyle olur.

Destek Vektör Makineleri eğitim verilerinde herhangi bir noktadan en uzak olan iki sınıf arasında bir karar sınırı bulan vektör uzayı tabanlı bir makine öğrenmesi yöntemidir. Bu yöntemle dağılım hakkında herhangi bir ön bilgi varsayımı yoktur, eğitim setlerinde girdi ve çıktılar eşlenir, eşler aracılığıyla test setinde ve yeni veri setlerinde değişkeni sınıflayacak karar fonksiyonları elde edilir.

Bu ödevde yukarıda bahsedilen iki yöntemin WEKA'da gerçeklenmeleri olan Multiple Perceptron ve SMO fonksiyonlarının başarıları karşılaştırılacaktır.

---

[1] https://tr.wikipedia.org/wiki/Weka

Uygulama

Bu ödevde Waikato Environment for Knowledge Analysis (WEKA) Version 3.8.1, *Multilayer Perceptron* ve *SMO* kütüphaneleri kullanılmıştır.

Sınıflandırma yapılacak veri seti http://yann.lecun.com/exdb/mnist/ adresinden edinilmiştir. Veri seti MNIST veritabanında bulunan el yazısı karakterlerden elde edilmiş 0-9 arası sayılardan oluşturulmuştur. Dosyadaki her satırdaki ilk 64 değer sayıysa ait özellikleri (gri seviyesinde piksel değerleri), 65. Değer ise sayının kaç olduğunu göstermektedir.

Ön İşlem

Veri seti WEKA ortamında açılmadan önce ortamın formatı olan ".arff"'ye dönüştürülmüştür.

Dosya özeti:

```
@RELATION numbers

@ATTRIBUTE A1           NUMERIC
@ATTRIBUTE A2           NUMERIC
.
.
@ATTRIBUTE A63          NUMERIC
@ATTRIBUTE A64          NUMERIC
@ATTRIBUTE class        {0,1,2,3,4,5,6,7,8,9}

@DATA
0,1,6,15,12,1,0,0,0,7,16,6,6,10,0,0,0,8,16,2,0,11,2,0,0,5,16,3,0,5,7,0,0,7,13,3,0,8,7,0,0,4,12,0,1,13,5,0,0,0,14,9,15,9,0,0,0,0,6,14,7,1,0,0,0
0,0,10,16,6,0,0,0,0,7,16,8,16,5,0,0,0,11,16,0,6,14,3,0,0,12,12,0,0,11,11,0,0,12,12,0,0,8,12,0,0,7,15,1,0,13,11,0,0,0,16,8,10,15,3,0,0,0,10,16,15,3,0,0,0
.
.
0,0,6,16,2,0,0,0,0,0,15,10,0,0,0,0,0,6,16,3,0,0,0,0,0,9,14,0,0,0,0,0,0,12,13,11,12,12,3,0,0,7,16,15,12,13,13,0,0,2,15,12,2,8,15,0,0,0,5,16,16,16,5,0,6
0,0,2,15,16,13,1,0,0,0,3,7,10,16,10,0,0,0,0,0,0,11,11,0,0,0,0,2,8,15,5,0,0,0,0,9,16,16,8,0,0,0,0,2,16,5,0,0,0,0,0,12,7,0,0,0,0,0,4,14,1,0,0,0,7
%
```

Sınıflandırma

Sınıflandırma yapılırken Eğitim ve Test seti olarak veri seti iki farklı şekilde ayrılmıştır; Cross-validation with Folds ve Percentage Split %85.

Multilayer Perceptron ve SMO kütüphaneleri için kullanılan parametreler aşağıdaki gibidir:

```
weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
```

```
weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K
"weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator
"weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4"
```

Veri Seti Hakkında Özet Bilgi

0-9 arasındaki sayıları 64 pikselin gri seviyeleri ile ifade eden veri setinde 3823 örnek bulunmaktadır. Veri seti rasgele bir şekilde %50 öğrenme ve %50 test etme verisi olacak şekilde ayrılmıştır. Dosyadaki ilk 64 veri attribute ve 65. veri sınıf olarak belirlenmiştir.

10 Fold Cross Validation

Verileri 10 Fold Cross Validation yöntemi ile ayrılması durumunda sonuçlar her iki yöntem için aşağıdaki şekilde olmuştur.

10 Fold Cross Validation	YSA		DVM	
Doğru Sınıflandırılmış Örnekler	3756	98,25%	3752	98,14%
Yanlış Sınıflandırılmış Örnekler	67	1,75%	71	1,86%
Ortalama Mutlak Hata	0,0052		0,1601	
Toplam Örnek Sayısı	3823		3823	

Percentage Split %85

Verilerin %85 eğitim, %15 test seti şeklinde ayrılması durumunda sonuçlar her iki yöntem için aşağıdaki şekilde olmuştur.

Test seti 85%	YSA		DVM	
Doğru Sınıflandırılmış Örnekler	559	97,56%	564	98,43%
Yanlış Sınıflandırılmış Örnekler	14	2,44%	9	1,57%
Ortalama Mutlak Hata	0,0058		0,1602	
Toplam Örnek Sayısı	573		573	

Sonuçların Yorumlanması

Her iki veri ayrımı senaryosu ve öğrenme algoritmasında elde edilen sonuçlar birbirine oldukça yakındır. Cross validation yapılmış senaryoda YSA %0,1 daha başarılıyken %15 test

verisi durumunda DVM %0,87 daha başarılı olmuştur. Tanıma başarısı açısında iki yöntemden biri daha başarılıdır demek mümkün olmamıştır.

YSA

Etiket \ Model	0	1	2	3	4	5	6	7	8	9	
	58	0	0	0	1	0	1	0	0	0	**0**
	0	56	0	0	0	0	0	0	0	0	**1**
	0	0	52	0	0	0	0	0	0	0	**2**
	0	0	0	56	0	0	0	1	0	0	**3**
	0	0	0	0	48	0	0	0	0	0	**4**
	0	0	0	0	0	59	0	0	0	3	**5**
	0	0	0	0	0	0	62	0	0	0	**6**
	1	0	0	0	1	0	0	46	0	0	**7**
	0	1	0	1	1	0	0	0	58	0	**8**
	0	0	0	1	1	1	0	0	0	64	**9**

DVM

Etiket \ Model	0	1	2	3	4	5	6	7	8	9	
	59	0	0	0	0	0	1	0	0	0	**0**
	0	56	0	0	0	0	0	0	0	0	**1**
	0	1	51	0	0	0	0	0	0	0	**2**
	0	0	0	56	0	0	0	1	0	0	**3**
	0	0	0	0	48	0	0	0	0	0	**4**
	0	0	0	0	0	61	0	0	0	1	**5**
	0	0	0	0	0	0	62	0	0	0	**6**
	0	0	0	0	0	0	0	48	0	0	**7**
	0	2	0	0	1	0	0	0	58	0	**8**
	0	0	0	1	0	1	0	0	0	65	**9**

%85 öğrenme durumu için verilen karışıklık matrisleri kıyaslandığında her iki model için de özellikle 8 ve 9 sayılarının daha çok yanlış tanındığı görülmüştür.

İki yöntem süre açısından kıyaslandığında öğrenme süresi açısından DVM öne çıkmaktadır.

	Süre (sn)			
	85% + %15		**Cross Validation**	
	YSA	**DVM**	**YSA**	**DVM**
Model	77,04	0,49	76,11	1,14
Test	0,02	0,01		
Toplam	77,06	0,5	76,11	1,14

Sonuç

Bu ödevde el yazısı tanıma için iki farklı makine öğrenmesi yöntemi (Yapay Nöron Ağları ve Destek Vektör Makineleri) WEKA aracı kullanılarak denenmiştir. Tanıma başarıları arasında büyük bir farklılık gözlemlenmemiştir. İki yöntem için de başarı yaklaşık olarak %98 seviyelerindedir. Öğrenme süreleri göz önüne alındığında Destek Vektör Makineleri açık ara öne geçmektedir. Veri seti Destek Vektör Makineleri kullanmaya müsait olduğu durumlarda Yapay Nöron Ağları yöntemine tercih edilebileceği söylenebilir.