Metin Sınıflandırma için Yeni Bir Eğitimli Anlamsal Özellik Seçimi Yöntemi

Melike TUTKAN Murat Can GANİZ Selim AKYOKUŞ

Bilgisayar Mühendisliği Bölümü
Mühendislik Fakültesi
Doğuş Üniversitesi, Kadıköy, İSTANBUL

E-mail: {mtutkan, mcganiz, sakyokus}@dogus.edu.tr

Özet

Metin sınıflandırmasında veri ön işleme metotlarının en önemlilerinden biri de özellik seçimidir. Özellik seçimi metin sınıflandırıcılarının doğruluğunu ölçeklenebilirliğini ve performansını iyileştirmektedir. Önerilen özellik seçimi algoritması, Helmholtz prensibi tabanlı Gestalt teorisine dayanan anlam isimli bir yöntemi temel almaktadır. Bu yöntem daha önce doküman özetleme ve özellik çıkarımı için kullanılmıştır. Bu çalışmada kelimelerin her bir sınıf için anlamının hesaplandığı ve bunların genel bir sıralama için birleştirildiği bir özellik seçimi yöntemi önerilmektedir.

1. Giriş

Metin madenciliğinde makine öğrenmesi algoritmalarıyla sınıflandırma yaparken en büyük sorunlardan birisi, özellik olarak kelimeler kullanıldığında özellik boyutunun çok yüksek olmasıdır. Gerçek hayatta böyle veri kümeleri çoğunluktadır ve genellikle bu özelliklerin bir kısmı gürültü içerir. Bunlar üzerinde çalışma yapmak için verideki sınıflandırma açısından etkin özelliklerin seçilmesi ve verinin temizlenmesi önemli ön adımlardan birisidir. Eğer kirli, etkinliği az ve çok sayıda özellik içeren bir veri kümesi üzerinde metin sınıflandırma yapılırsa ortaya çıkan sonuç tutarlı ve güvenilir olmayabilir. Özellik seçimi metotları, veri kümesinin sınıflandırılması için daha önemli olan özelliklerin seçimini sağlar. Bundan dolayı, yapılan metin sınıflandırma uygulamaları daha hızlı çalışırken daha etkin sonuçlar üretebilir. Özellik seçimi konusunda birçok yayın yapılmıştır [1][2] ve bu konu hala önemli araştırma alanlarından bir tanesidir [3].

Bu çalışmada eğitimli anlamsal özellik seçimi isimli yeni bir özellik seçimi metodu geliştirilmiştir. Daha önceki çalışmalarda anlam değeri doküman özetleme [4] ve özellik çıkarımı [5] için kullanılmış ve daha çok cümleler ve paragraflar üzerinde çalışılmıştır.

Bu çalışmada farklı veri kümelerindeki sınıflandırma açısından etkin özellikler anlam değeri yöntemiyle belirlenmiştir. Elde edilen etkin özellikleri içeren veri kümeleri üzerinde WEKA isimli makine öğrenmesi araç kiti [6] kullanılarak özellik seçim yöntemlerinin sınıflandırma başarımına etkisi ölçülmüştür. Sınıflandırıcı olarak Multinomial Naive Bayes (MNB) seçilmiştir. Bunun sebebi MNB'nin metin sınıflandırmada çok kullanılan etkili ve hızlı bir sınıflandırıcı olması ve özellik seçimi yöntemlerine yüksek duyarlılık göstermesidir.

Bu bildiri aşağıda anlatıldığı şekilde düzenlenmiştir: ikinci bölümde anlam değerinin çalışmamızdaki tanımı yapılmış; üçüncü bölümün ilk kısmında anlam değerinin eğitimli anlamsal özellik seçimi metoduna nasıl dönüştürdüğümüz açıklanmış, üçüncü bölümün ikinci kısmında ise deneylerde kullandığımız diğer özellik seçimi metotları anlatılmış; dördüncü bölümünde kullanılan veri kümeleri tanıtılmış; beşinci bölümün ilk kısmında sınıf tabanlı olarak hesaplanan anlam değeri yüksek olan her sınıfa ait ilk 10 kelime sunulmuş, beşinci bölümün ikinci kısmında ise önerilen eğitimli anlamsal özellik seçimi metodu ile başka özellik seçimi metotları MNB sınıflandırma algoritmasının doğruluk değerlerine etkisi ölçülmüş, son bölümde ise sonuçlar üzerinde durulmuştur.

2.Anlam Değeri

Anlam değeri Helmholtz prensibi tabanlı Gestalt teorisine dayanmaktadır. Bu çalışmada anlam değeri literatürde ilk defa bir eğitimli özellik seçimi yöntemi olarak kullanılmıştır. Anlam değeri aşağıdaki formüllerle hesaplanmaktadır.

$$Anlam(k,P,D) = -\frac{1}{m}\log YAS(k,P,D) \quad (1)$$

$$YAS(k,P,D) = \binom{K}{m}\frac{1}{N^{m-1}} \quad (2)$$

Anlam değeri formülü daha önceki çalışmalarda doküman özetleme [4] ve özellik çıkarımı [5] yöntemlerinde her dokümanın (*D*), içindeki paragraf ve cümle gibi parçalarının (*P*), içerdiği kelimelerin (*k*) anlam değerini hesaplamak için kullanılmıştır. Anlam değerinin yüksek olması kelimenin daha önemli olduğunu göstermektedir. Bu çalışmada aynı formülü eğitimli özellik seçimi için kullandık. Çalışmamızda tüm veri kümesini bir doküman (*D*), veri kümesi içerisinde yer alan farklı sınıfları veri kümesinin parçaları (*P*), veri kümesindeki özellikleri ise kelimeler (*k*), olarak kabul ettik. Bu varsayıma bağlı olarak formülde yer alan her terimin açıklaması aşağıda verilmiştir:

k: özellik (kök kelime, terim)
P: bir sınıfa ait dokümanlar
D: veri kümesindeki tüm dokümanlar
m: bir sınıf içerisinde bulunan dokümanlarda *k* özelliğinin geçme sayısı
K: tüm veri kümesinde *k* özelliğinin geçme sayısı
N: tüm dokümanların uzunluğunun (toplam kelime sayısı) bir sınıfa ait dokümanların uzunluğuna (toplam kelime sayısı) bölümüdür.

$$N = \frac{|D|}{|P|} \quad (3)$$

Anlam değerinin hesaplanmasında kullanılan YAS (Yanlış Alarm Sayısı) değeri, anlam değeri ile ters orantılıdır. Bunun anlamı YAS değeri ne kadar küçükse o özelliğin o sınıf için anlam değeri o kadar büyüktür. Anlam değerinin büyük olması ise özelliklerin daha etkin ve önemli bir özellik olduğunu ifade etmektedir.

3. Özellik Seçimi Metotları

3.1. En Büyük Anlamsal Özellik Seçimi (EAÖS)

İkinci bölümde açıklanan anlam formülü ile her bir özelliğin her bir sınıftaki dokümanlar için anlam değeri hesaplanmaktadır. Her bir özellik için sınıf sayısı (|s|) kadar anlam değeri hesaplanmaktadır. En iyi özelliği seçmek için aşağıdaki yaklaşım kullanılmıştır.

Her bir özellik için ayrı ayrı, sınıf tabanlı olarak hesaplanan anlam değerlerine bakılarak, en yüksek anlam değeri o özelliğin anlam değeri olarak kabul edilmiş ve bu metoda "EAÖS" (Eğitimli En Büyük Anlamsal Özellik Seçimi) ismi verilmiştir.

Bu yaklaşımla listenin üst kısmında önemli ve etkin özellikler, alt kısmında ise daha az etkin ve önemsiz özellikleri içerecek şekilde sıralanmıştır. Belirlenen özellik azaltma oranına göre listenin başından sırasıyla en iyi özellikler seçilerek özellik seçimi için kullanılabilmektedir.

3. 2. EOR, MOR, WOR, CDM

Bu çalışmada anlam formülüne dayanan yeni önerilmiş olan EAÖS (Eğitimli En Büyük Anlamsal Özellik Seçimi) yöntemi; EOR, MOR, WOR, CDM olarak kısaltılan ve aşağıda açıklanan özellik seçimi yöntemleri karşılaştırılmıştır. Bu yöntemler, klasik göreceli olasılıklar oranı (Odds Ratio) formülünün üzerinde çalışarak ve bir takım değişiklikler yapılarak ortaya atılan özellik seçimi metotlarıdır ve (4),(5),(6) ve (7)'daki formüllerle hesaplanmaktadır [2].

EOR (Extended Odds Ratio: Genişletilmiş Göreceli Olasılıklar Oranı):

$$EOR(w) = \sum_j \frac{\log P(w|c_j)(1-P(w|\overline{c}_j))}{\log P(w|\overline{c}_j)(1-P(w|c_j))} \quad (4)$$

WOR (Weighted Odds Ratio: Ağırlıklandırılmış Göreceli Olasılıklar Oranı):

$$WOR(w) = \sum_j P(c_j)\frac{\log P(w|c_j)(1-P(w|\overline{c}_j))}{\log P(w|\overline{c}_j)(1-P(w|c_j))} \quad (5)$$

CDM (Class Discriminating Measure: Sınıf Ayırıcı Ölçü):

$$CDM(w) = \sum_j \left|\frac{\log P(w|c_j)}{\log P(w|\overline{c}_j)}\right| \quad (6)$$

MOR (Multi-Class Odd Ratio: Çok Sınıflı Göreceli Olasılıklar Oranı):

$$MOR(w) = \sum_j \left|\frac{\log P(w|c_j)(1-P(w|\overline{c}_j))}{\log P(w|\overline{c}_j)(1-P(w|c_j))}\right| \quad (7)$$

Formüllerdeki $P(w|c_j)$ ifadesi *w* kelimesinin c_j sınıfında olma olasılığını, $P(w|\overline{c_j})$ ifadesi ise *w* kelimesinin c_j sınıfı hariç diğer tüm sınıflarda olma olasılığını göstermektedir.

4. Veri Kümeleri

Bu çalışmada iki farklı Türkçe veri kümesi kullanıldı. Bu kümelerdeki sınıf, doküman ve özellik sayıları Tablo1'de sunulmuştur. Veri kümelerinde deneme amaçlı olarak farklı fakat benzer performansa sahip kök indirgeme yöntemleri kullanılmıştır.

1150haber veri kümesi, 2004 yılında hazırlanan 5 farklı haber sınıfına ait 230'ar toplamda 1150 haber içermektedir [8]. Haber metninin sınıfları: Ekonomi, Magazin, Sağlık, Siyasi ve Spor'dur. FPS5 [9] adlı kök indirgeme yöntemi kullanılarak özellikler en fazla 5 karakter olacak şekilde köklerine ayrılmıştır. Milliyet4c1k veri kümesi, 4 farklı haber sınıfına ait 2002-2011 yılları arasında toplanmış 1000'er toplamda 4000 adet haber içermektedir. Haber metninin sınıfları: Dünya, Ekonomi, Siyaset ve Spor'dur [10]. Zemberek kullanılarak özellikler köklerine indirgenmiştir [11]

Tablo 1.Veri kümelerinin boyutları

Veri kümesi	|Sınıf|	|Doküman|	|Özellik|
1150haber	5	1.150	6.656
Milliyet4c1k	4	4.000	21.469

5. Bulgular

5. 1. Anlam Değeri Yüksek Özellikler

Sınıf tabanlı olarak hesaplanan anlam hesaplamaları sonucunda 1150haber veri kümesindeki her sınıf için en önemli ilk 10 özellik Tablo 2'de, Milliyet4c1k veri kümesindeki her sınıf için en önemli ilk 10 özellik Tablo 3'te sunulmuştur.

Tablo 2. 1150haber veri kümesi için en yüksek değere sahip ilk 10 özellik (kök kelime)

Ekonomi	**Magazin**	**Sağlık**	**Siyasi**	**Spor**
cari	pekin	tümör	anaya	maçta
borsa	hande	ultra	annan	luces
açığı	pekka	ışınl	kerkü	sahad
döviz	sosye	cildi	dgm	orteg
varil	ataiz	lazer	aihm	stadı
unakı	madon	kanam	mhp	dk
tahvi	laila	enfek	mgk	tribü
mevdu	ajda	menop	laikl	defan
ötv	dizid	cilt	bayar	golle
venez	çapkı	kasla	şaron	depla

Bu veri kümelerindeki haberler incelendiğinde veri kümesinin toplandığı tarihlerdeki önemli kavram ve isimlerin listenin başında yer aldığı gözükmektedir.

Tablo 3. Milliyet4c1k veri kümesi için en yüksek değere sahip ilk 10 özellik (kök kelime)

Dünya	**Ekonomi**	**Siyaset**	**Spor**
osetya	gdo	balbay	hiddink
abhazya	gsyh	karabekir	forvet
saakaşvil	kobi	tutukluluk	ankaragüç
sarko	zeytinyak	kanadoğlu	bobo
pelos	goldma	selek	mhk
ateşke	fed	oramiral	gs
muharip	likidi	yarsav	tff
ayetullah	sachs	johannesburgu	stoper
domino	perake	tatar	faul
annapolis	karlılık	fige	lacivertli

5.2. Özellik Seçimi Yöntemlerinin Karşılaştırılması

Eğitimli anlamsal en büyük özellik seçimi "EAÖS" metodunun sınıflandırma performansına etkisini ölçebilmek için daha önceki çalışmalarda ortaya atılan ve iyi performans gösterdiği iddia edilen EOR, MOR, WOR ve CDM [2] özellik seçimi metotlarıyla karşılaştırılması bu bölümde yapılmıştır.

Özellik seçimi metotlarıyla özellik sayısını azalttıktan sonra MNB sınıflandırıcısının doğruluk başarımı 10 kat çapraz doğrulama (10-fold Cross Validation) yöntemi ile ölçülmüştür.

Tablo 4, 1150haber veri kümesinde kullanılan özellik seçimi yöntemlerinin MNB'in sınıflandırma performansına etkisini göstermektedir. Özellik seçimi yöntemleri kullanılarak en önemli 500, 1000, 2000, 3000, 4000, 5000, 6000 özellik seçilmiş bu sayı (|Ö|) "Milliyet4c1k" için 1000'er artarak 10.000'e kadar devam etmektedir. Tablolardaki değerler MNB sınıflandırıcısının doğruluk değerlerini içermektedir. Tablo 5 aynı şekilde Milliyet4c1k veri kümesi için yapılmış olan deney sonuçlarını göstermektedir.

Tablo 4. 1150 haber veri kümesi farklı özellik seçimi metotlarıyla MNB sınıflandırma doğruluğu

|Ö|	EAÖS	EOR	CDM	MOR	WOR
500	75,04	62,17	92,61	**92,70**	62,35
1000	86,26	71,65	93,74	**93,83**	71,57
2000	92,35	79,57	**93,83**	**93,83**	79,57
3000	**94,17**	83,39	93,74	93,83	83,39
4000	**94,96**	87,13	93,83	93,83	87,13
5000	**94,61**	89,57	93,91	93,91	89,57
6000	**94,35**	91,91	94,09	94,09	91,91

Tablo 5. Milliyet4c1k veri kümesi farklı özellik seçimi metotlarıyla MNB sınıflandırma doğruluğu

|Ö|	EAÖS	EOR	CDM	MOR	WOR
500	60,78	67,00	81,88	**81,93**	67,00
1000	67,40	66,33	**85,08**	**85,08**	66,33
2000	76,70	77,18	**86,35**	**86,35**	77,18
3000	81,78	80,15	**87,00**	**87,00**	80,15
4000	84,10	81,85	**87,13**	87,10	81,85
5000	**87,60**	81,40	87,33	87,35	81,40
6000	**88,18**	82,45	87,43	87,43	82,45
7000	**89,38**	83,25	87,90	87,90	83,25
8000	**88,78**	83,75	87,88	87,88	83,75
9000	**89,10**	83,78	88,00	88,03	83,78
10000	**88,78**	84,53	88,23	88,25	84,53

Tablo 4'ün başarı değerleri analiz edildiğinde EAÖS, EOR ve WOR'dan genel olarak daha iyi sonuçlar verdiği ayrıca seçilen özellik sayısı arttıkça EAÖS başarımının CDM ve MOR'un değerlerini geçtiği gözlemlenmektedir. Tablo5'in başarı değerleri analiz edildiğinde EAÖS yönteminin 4000'e kadar olan özellikler sayılarında diğer özellik seçimi yöntemleri kadar iyi çalışmadığı fakat seçilen özellik sayısı arttıkça ilk başta EOR ve WOR yöntemlerini, daha sonra her iki veri kümesinde de iyi sonuçlar veren CDM ve MOR'un başarımını geçtiği görülmektedir. Burada en önemli gözlemlerden birisi her iki veri kümesinde de en yüksek doğruluk değerlerinin EAÖS yöntemi ile elde edildiğidir.

6. Sonuç

Bu çalışmada anlam değeri formülü farklı bir alan olan özellik seçimi alanında kullanılarak yeni bir eğitimli özellik seçimi yöntemi olarak sunulmuştur. Sonuçlar incelendiğinde yeni sunulan EAÖS metodunun iyi başarım veren diğer metotlara kıyasla, seçilen özellik sayısı arttıkça, başarımının daha fazla olduğu gözlemlenmektedir. Ayrıca her iki veri kümesinde de en yüksek doğruluk değerleri EAÖS yöntemi ile elde edilmiştir.

Gelecek çalışmalarda yeni yöntemi farklı veri kümelerinde, farklı sınıflandırıcılarla birlikte denemeyi ve özellik seçimi yöntemleri arasında oldukça yaygın kullanılan Bilgi Kazancı (IG-Information Gain) ve Ki-kare (Chi-squared) yöntemleri ile de karşılaştırmayı planlamaktayız. Ayrıca en iyi özellikleri seçmek için bu bildiride kullandığımız "en büyük" işlemi yerine başka metotlar üzerinde de çalışmaktayız.

7. Teşekkür

Bu çalışma TÜBİTAK tarafından 111e239 no'lu proje ile kısmi olarak desteklenmiştir.

8. Kaynaklar

[1] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, Z. Wang, "A novel feature selection algorithm for text categorization", Expert Systems with Applications 33,1-5,2007

[2] J. Chen, H. Huang, S. Tian, Y. Qu, "Feature selection for text classification with Naïve Bayes", Expert Systems with Applications 36, 5432-5435, 2009

[3] S. Baccianella, A. Esuli, F. Sebastiani, "Using micro-documents for feature selection:The case of ordinal text classification", Expert Systems with Applications 40, 4687-4696, 2013

[4] H. Balinsky, A. Balinsky, and S. Simske, "Document sentences as a small world," Proc. of IEEE SMC, 2011

[5] A. Balinsky, H. Balinsky, and S. Simske, "On the Helmholtz principle for data mining," Proc. of 2011 Conf. on Knowledge Discovery, Chengdu, China, April 2011

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H, Witten The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1, 2009

[7] A. Desolneux, L. Moisan, and J.-M. Morel, From Gestalt Theory to Image Analysis: A Probabilistic Approach, ser. Interdisciplinary Applied Mathematics, Springer, vol. 34, 2008

[8] M.F. Amasyalı, A. Beken, "Türkçe Kelimelerin Anlamsal Benzerliklerinin ölçülmesi ve Metin sınıflandırmada Kullanılması",Sui Antalya, 2009

[9] F. Can, S. Kocberber, E. Balcık, C. Kaynak, H.C. Ocalan, O. M. Vursavas, "Information Retrieval On Turkidh Texts", Journal of the American Society For Information Sceince andTechnology, Vol,59, No,3, Pp, 407-421, February 2008

[10] M. Poyraz, Z.H. Kilimci, M.C. Ganiz, (2014). Higher-Order Smoothing: A Novel Semantic Smoothing Method for Text Classification. Journal Of Computer Science and Technology , Vol.29, No.3, 2014, pp.376-391

[11] A.A. Akın, M.D. Akın. "Zemberek, an open source NLP framework for Turkic Languages." *Structure* 10 (2007).