Metin Sınıflandırma için Eğitimsiz bir Anlamsal Özellik Seçimi Yöntemi
An Unsupervised Semantic Attribute Selection Method for Text Classification

Melike TUTKAN, Murat Can GANİZ, Selim AKYOKUŞ

Mühendislik Fakültesi
Bilgisayar Mühendisliği Bölümü
Doğuş Üniversitesi
{mtutkan, mcganiz, sakyokus}@dogus.edu.tr

Özet

Metin sınıflandırma için önemli sorunlardan birisi yüksek boyutlu özellik uzayıdır. Özellik seçimi, metin sınıflandırma için kullanılan veri ön işleme metotlarından birisidir. Özellik seçimi özellik sayısını azaltarak metin sınıflandırıcısının etkinliğini arttırır. Bu çalışma Helmholtz prensibi temelli Gestalt teorisine dayanan anlam formülü hesabına dayanmaktadır ve bu çalışmada anlam formülünü kullanarak yeni bir eğitimsiz anlamsal özellik seçimi metodu önerdik. Bu çalışmanın amacı, anlam değerini özellik seçimi olarak uygulayarak yüksek boyutlu özellik uzayının veri boyutunu azaltmaktır. Böylelikle sınıflandırıcıların sınıflandırma için harcadığı zaman azalalırken aynı zamanda karmaşıklığı daha az olan modeller üretilebilir. Daha önceki çalışmalarda anlam formülü, doküman özetleme ve özellik çıkarımı için kullanılmıştır fakat bu çalışma aynı yöntem eğitimsiz özellik seçimi için kullanılmaktadır.

Abstract

For text classification a one of the major problem is the high dimensionality of the feature space. Feature selection is one of the preprocessing methods for text classification. Feature selection reduce number of feature so it improves efficiency of text classifiers. Our study is based on meaning formula which is based on Helmholtz principle from the Gestalt theory and we propose a new Unsupervised Meaning Feature Selection method with using meaning formula. Purpose of this study is to apply meaning measure for feature selection in order to decrease the number of features of high dimentional feature space. So that the time spent for classification can be decreased and less complex models can be produced. In previous studies, meaning formula is used for document summarization and feature extraction however in this study same method is used for unsupervised feature selection.

1. Giriş

Kelime sepeti yaklaşımı ile metinleri temsil için kelimeler veya terimler kullanıldığında on binlerce farklı özellik ortaya çıkabilir fakat bu özelliklerin tümü sınıflandırma için ilgili ve yararlı olmayabilir. Özellik olarak alınan fakat gerçekte özellik olabilecek kadar etkinliği olmayan -ki bunlara gürültü diyoruz- özellikler sınıflandırma doğruluğunu düşürebilir. Ayrıca yüksek boyutlu özellik uzayı sınıflandırıcının sınıflandırma için harcadığı zamanı ve hafızayı arttırabilir. Bu problemi çözmek için bazı veri ön işleme metotları kullanılmaktadır. Biz bu bildiride özellik seçimi metodunu ele almaktayız. Özellik seçimi ile metin sınıflandırma için ilgili ve yararlı özelliklerden, ilgisiz ve sınıflandırma için pek bir yarar sağlamayan özellikleri ayırabiliriz. Küçülen özellik uzayında bulunan düşük sayıda fakat sınıflandırma için daha çok öneme sahip ve etkin özelliklerle yapılan sınıflandırmalar daha hızlı ve daha etkin sonuçlar çıkartabilir.

Özellik seçimi, üzerinde hala çalışılan bir veri ön işleme metodudur. Daha önceki yıllarda yazılmış bir yayında [1] karar ağaçlarının kullandığı "Gini index"in özellik seçimi olarak kullanılabileceği, başka bir yayında [2] ise beş yeni özellik seçimi metriği sunulmuştur. Son yıllarda yazılan yayın [3] ise bu konunun güncelliğini koruduğunu göstermektedir. Bu çalışmada ise eğitimsiz ve anlam değerini kullanan yeni bir özellik seçimi sunulmuştur.

Anlam değeri Helmholtz prensibi tabanlı Gestalt teorisine dayanan matematiksel bir formüldür. Doküman özetleme [4] ve özellik çıkarımı [5] için kullanılan bu formül küçük yapıda doküman parçaları üzerinde denenmiştir. Biz çalışmamızda bu formülü eğitimsiz anlamsal özellik seçimi (EAÖS) için kullanmak üzere düzenledik. Deneylerimizde önerilen EAÖS yönteminin sınıflandırma doğruluğuna etkisi ölçülmüştür. Yöntem literatürde sık olarak kullanılan eğitimli Bilgi Kazancı (BK)[10] ve Ki-kare (χ^2)[11] yöntemleri ile kıyaslanmış ve bazı şartlarda benzer veya daha iyi çalıştığı gösterilmiştir.

Deneyleri yaparken WEKA isimli Waikato üniversitesinde geliştirilmiş makine öğrenmesi araç kiti kullandık [6]. WEKA'da makine öğrenimi algoritmaları ve metotları hazır olarak bulunmakta ve bu konu üzerinde araştırma yapan kişiler için ücretsiz dağıtılmaktadır.

Bildiri aşağıda anlatıldığı gibi düzenlenmiştir: 2. bölümde anlam değerinin çalışmamızdaki kullanımının tanımı yapılmış, 3. bölümde anlam değerinin eğitimsiz anlamsal özellik seçimi metoduna dönüşümünü açıklamış, 4. bölümde kullanılan veri kümeleri tanıtılmış, 5. bölümün ilk kısmında tüm veri kümesi için anlam değeri en büyük olan ilk 10 kelime ile diğer özellik

seçimi metotlarının çıkardığı en yüksek değere sahip özellikler çizelge halinde sunulmuş, 2. kısmında ise önerilen özellik seçimi metodu ile yaygın kullanılan diğer özellik seçimi metotlarının indirgediği özellik uzaylarıyla çalıştırılan Multinomial Naive Bayes (MNB) sınıflandırıcısının doğruluk değerleri karşılaştırılmış ve son bölümde sonuçlar tartışılmıştır.

2. Anlam Değeri

Helmholtz prensibi tabanlı Gestalt teorisine [5] dayanan anlam değeri (1), (2) ve (3)'deki formüllerle hesaplanmaktadır.

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

$$Anlam(k,P,D) = -\frac{1}{m} \log YAS(k,P,D) \quad (2)$$

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

Anlam değeri formülü ile *D*'nin bir parçası olan *P*'nin içerisindeki *k* teriminin anlam değeri hesaplanmaktadır. [4]'te *D,* dokümanın paragrafları, *P* ise o paragraflardaki cümle olarak belirlenmiş ve ona göre anlam değerleri hesaplanmıştır. Fakat biz bu formülü kullanırken *D*'yi veri kümesinin tümü, *P*'yi ise o veri kümesindeki doküman olarak aldık. Bu varsayım ile formülde yer alan her bir terimin açıklaması aşağıda verilmiştir:

k: özellik (kök kelime)
P: veri kümesindeki dokümanların her biri
D: veri kümesindeki tüm dokümanlar
m: bir doküman içerisinde *k* özelliğinin geçme sayısı
K: tüm veri kümesinde *k* özelliğinin geçme sayısı
N: tüm veri kümesinin boyunun (toplam kelime sayısı), bir dokümanın boyuna (toplam kelime sayısı) bölümü

Burda doküman sayısına *d* dersek, her bir özelliğin *d* adet anlam değeri oluşmaktadır. Anlam değeri için hesaplanan yanlış alarm sayısı (YAS) değeri, anlam değeri ile ters orantılıdır. Başka bir deyişle YAS değeri ne kadar ufak çıkarsa o özelliğin anlam değeri o kadar büyük olur. Bu sonuç ise bize bu özelliğin daha etkin ve önemli bir özellik olduğunu gösterir.

Daha önceki çalışmalarda anlam formülü çok az terim içeren düşük boyuttaki dokümanlar için kullanıldığından çalışmamızda kullandığımız büyük boyuttaki dokümanların hesaplamalarında programlamada kullandığımız değişkenlerde taşma meydana gelmiştir. Bu problemin çözümü için bazı matematiksel dönüşüm ve hesaplama yöntemleri kullanılmış, anlam formülü aşağıdaki aşamalardan geçerek en son (6)'daki halini almıştır.

$$\log YAS(k,P,D) = \log\left(\binom{K}{m} \frac{1}{N^{m-1}}\right) \quad (4)$$

$$\log YAS(k,P,D) = \log\left(\binom{K}{m}\right) + \log\left(\frac{1}{N^{m-1}}\right) \quad (5)$$

$$\log YAS\ (k,P,D) = \log\left(\binom{K}{m}\right) - ((m-1)\log(N) \quad (6)$$

3. Eğitimsiz Anlamsal Özellik Seçimi

Anlam formülü ile her bir özelliğin her bir doküman için anlam değeri hesaplanmaktadır. Bu durumdan dolayı her özelliğin doküman sayısı kadar anlam değeri oluşmaktadır. Sınıflandırmada kullanılacak en iyi özellikleri seçmek için aşağıdaki yaklaşım kullanılmıştır.

Her bir parça (*P*) için ayrı ayrı hesaplanan özelliklerin (*k*) anlam değerlerinin en büyüğü o özelliğin anlamsal değeri olarak kabul edilmiş bu metoda "EAEBÖS" (Eğitimsiz Anlamsal En Büyük Özellik Seçimi) ismi verilmiştir.

$$Anlam(k,D) = Enbüyük\,(Anlam(k,P,D)) \quad (7)$$

Bu yaklaşımla elde edilen anlam değeri listesi, anlam değeri yüksek olan başta olmak üzere büyükten küçüğe doğru özellik seçimi uygulamasında kullanılmak üzere sıralanmıştır.

4. Veri Kümeleri

Bu çalışmada eğitimsiz anlamsal özellik seçimi metodunun etkinliğini farklı dillerde görebilmek için bir İngilizce ve bir Türkçe olmak üzere iki farklı veri kümesi kullanılmıştır. Veri kümelerinin sınıf, doküman ve özellik sayıları Çizelge 1'de verilmiştir.

1150haber veri kümesi beş sınıfı olan her bir sınıfında 230'ar toplamda 1150 adet haber bulunduran Türkçe bir veri kümesidir [7]. Özellikler üzerinde köklerine ayırma işlemi gerçekleştirilmiştir. Bu işlem özelliğin ilk 5 harfininin kök olarak alınması şeklinde yapılmıştır [8].

Tr31 veri kümesi yedi sınıfı, toplamda 927 dokümanı olan bir veri kümesidir [9]. Her bir sınıftaki doküman ve özellik sayıları Çizelge 1'de verilmiştir. Tr31 veri kümesinde her sınıfta farklı sayıda doküman bulunmakta olup bunların sayıları Çizelge 2'de verilmiştir.

Çizelge 1: Veri kümelerinin boyutları

Veri kümesi	Sınıf	Doküman	Özellik
1150haber	5	1,150	6,656
Tr31	7	927	10,129

Çizelge 2: Tr31 veri kümesinin sınıflar için doküman dağılımı

Sınıf	Doküman
0	352
1	63
2	151
3	21
4	227
5	111
6	2

5. Deney Sonuçları

5.1. Anlam Değeri Deney Sonuçları

Bu bölümde yeni sunulan özellik seçimi yöntemini ile yapılan deney sonuçları yer almaktadır. Anlam değeri hesaplamalarından sonra elde edilen her bir kelimenin anlam değeri artan sıra ile sıralanmış ve ilk 10 kelime, diğer özellik seçimi yöntemlerinin veri kümesi için hesapladığı en önemli ilk 10 kelime ile kıyaslanmıştır. 1150haber Türkçe veri kümesi için bu kıyaslama Çizelge 3'te, Tr31 İngilizce veri kümesi için ise Çizelge 4'te verilmiştir.

Çizelge 3: 1150haber veri kümesi için önemli ilk 10 özellik

Yeni Metot (EAEBÖS)	BK	χ^2
yonga	takım	takım
horla	futbo	futbo
farel	hasta	hasta
umre	maçın	maçın
köpeğ	yüzde	şampi
shui	oyunc	yüzde
feng	şampi	ankar
zerre	ankar	oyunc
çupi	dolar	dolar
berks	maç	direk

Çizelge 4: Tr31 veri kümesi için önemli ilk 10 özellik

Yeni Metot (EAEBÖS)	BK	χ^2
panda	drug	polio
inglewood	traffick	endang
matamoro	endang	speci
zachert	speci	drug
endesa	africa	traffick
estuari	cartel	hydroelectr
vladivostok	cocain	fountain
guinean	narcot	africa
kelantan	project	dam
fpr	wildlif	project

5.2. Anlamsal Özellik Seçimi Deney Sonuçları

Bu bölümde yeni sunulan özellik seçimi metodunun Multinomial Naive Bayes (MNB) algoritmasının sınıflandırma performansına etkisini görebilmek için yaygın olarak kullanılan özellik seçim metotları olan Bilgi Kazancı (BK)[10] ve Ki-kare(χ^2) [11]'i kullandık. Özellik seçimi metotlarıyla özellik sayısı indirgendikten sonra sınıflandırma performansını ölçtük. Buna ek olarak sınıflandırma doğruluğunu hesaplarken 10 kat çapraz doğrulama (10–fold cross-validation) yöntemini kullandık.

Çizelge 5, özellik seçimi yöntemleri uygulanarak özellik uzayı sırasıyla 500, 1000, 2000, 3000, 4000, 5000, 6000'e indirgenen 1150 haber veri kümesi üzerinde yapılan MNB sınıflandırıcısının elde ettiği sınıflandırma doğruluklarını göstermektedir. Aynı işlemler Tr31 veri kümesinde de yapılmış fakat bu veri kümesinin büyük boyutlu olmasından dolayı özellik uzayının sırasıyla 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000'e indirgenmiştir. Çizelge 6'da Tr31 veri kümesi için sınıflandırma sonuçları yer almaktadır.

Her bir özellik sayısı için elde edilen en yüksek sınıflandırma sonuçları kalın yazılarak belirginleştirilmiştir. Buradan her bir özellik sayısı seviyesinde hangi metodun daha iyi çalıştığı gözlemlenebilir.

Çizelge 5: 1150haber veri kümesinin farklı özellik seçimi metotlarıyla MNB sınıflandırma doğruluğu

|özellik|	Yeni Metot (EAEBÖS)	BK	χ^2
500	42.26	**92.87**	92.78
1000	68.00	**93.91**	93.65
2000	86.52	93.48	**93.65**
3000	92.70	**94.00**	93.91
4000	**94.00**	**94.00**	**94.00**
5000	93.74	**93.91**	**93.91**
6000	94.00	**94.26**	**94.26**

Çizelge 6: Tr31 veri kümesinin farklı özellik seçimi metotlarıyla MNB sınıflandırma doğruluğu

|özellik|	Yeni Metot (EAEBÖS)	BK	χ^2
500	81.55	90.51	**93.42**
1000	92.77	92.77	**93.20**
2000	**96.12**	94.07	93.85
3000	**95.90**	94.28	94.07
4000	**95.90**	94.39	94.39
5000	**95.69**	94.39	94.39
6000	**95.36**	94.39	94.39
7000	**94.93**	94.61	94.61
8000	**94.93**	94.71	94.71
9000	**94.71**	94.39	94.39
10000	**94.71**	94.61	94.61

Çizelgelerdeki başarım değerleri analiz edildiğinde anlamsal özellik seçimi metodunun, Türkçe veri kümesinde seçilen özellik sayısı arttıkça alınan doğruluk değerlerinin bazı durumlarda BK'ye yaklaştığı görülmekle birlikte kullanılan İngilizce veri kümesinde eğitimsiz anlamsal özellik seçimi metodunun oldukça iyi sonuçlar verdiği görülmektedir.

6. Sonuçlar

Bu çalışmada anlam değeri kullanılarak yeni bir özellik seçimi yöntemi tasarlanmıştır. Sonuçlar değerlendirildiğinde seçilen özellik sayısı arttıkça bu yöntemin başarısı diğer iki yaygın özellik seçimi yöntemi kadar, bazı durumlarda daha iyi sonuçlar verdiği gözlemlenmektedir.

Bunun yanında yeni sunulan özellik seçimi metodunun daha hızlı çalıştığı da gözlemlenmiştir. Deneylerde kullanılan bilgisayarın özellikleri şu şekildedir: Intel(R) Core(TM) i7-4770 CPU @ 3.40GHz, 16 GB RAM. Çizelge 7'te özellik

seçimi ve MNB'in yaptığı sınıflandırma işlemlerinin toplam harcadığı zaman saniye cinsinden verilmiştir.

Çizelge 7: MNB sınıflandırıcısının her bir özellik seçimi metodu ile toplam harcadığı zaman (saniye)

Veri kümesi	EAEBÖS	BK	χ^2
1150haber	27	365	366
Tr31	62	828	831

Eğitimsiz anlamsal özellik seçimi metodunu bundan sonraki çalışmalarda farklı veri kümeleri üzerinde ve farklı sınıflandırıcılarla denemeyi planlamaktayız, ayrıca 3. bölümde açıklanan "EAEBÖS" yaklaşımına ek olarak başka yaklaşımlar üzerinde de çalışmaktayız.

7. Teşekkür

Bu çalışma TÜBİTAK proje no: 111E239 tarafından kısmi olarak desteklenmiştir.

8. Kaynaklar

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

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

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

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

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

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1, 2009

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

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

[9] Han, E. and Karypis, G. "Centroid-Based Document Classification: Analysis & Experimental Results". In Proc. of the 4th European Conf. on the Principles of Data Mining and Knowledge Discovery (PKDD): 424-431, 2000.

[10] Tom Mitchell, Machine Learning McCraw Hill, 1996

[11] Yang Y., Jan O. Pedersen A Comparative Study on Feature Selection in Text Categorization, Proceedings of the Fourteenth International Conference on Machine Learning, p.412-420, July 08-12, 1997