SINIF BİLGİSİNİ KULLANAN BOYUT İNDİRGEME YÖNTEMLERİNİN METİN SINIFLANDIRMADAKİ ETKİLERİNİN KARŞILAŞTIRILMASI

COMPARING THE IMPACTS OF DIMENSION REDUCTION METHODS THAT USE CLASS LABELS ON TEXT CLASSIFICATION

Göksel Biricik

Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
goksel@ce.yildiz.edu.tr

ÖZETÇE

Örnekleri çok sayıda öznitelik barındıran veri kümelerinin sınıflandırılması uzay ve zaman olarak yüksek maliyetlidir. Çok boyutluluğun laneti olarak bilinen bu problemi çözmek için öznitelik seçimi ve öznitelik çıkarımı yöntemlerinden oluşan boyut indirgeme yöntemleri geliştirilmiştir. Bu çalışmada, her bir örnekte bulunan özniteliklerin sınıflara olan bileşke etkilerini kullanarak boyut indirgeme sağlayan soyut öznitelik çıkarım yöntemi ile sınıf bilgisini kullanarak boyut indirgemeyi gerçekleştiren diğer yaygın yöntemlerin sınıflandırma performansına olan etkileri karşılaştırılmıştır. Yöntemleri karşılaştırmak için çok boyutlu özniteliklerıyle bilinen iki standart metin veri kümesi kullanılmıştır. Seçilen yöntemlerle örneklerinin boyutları indirgenen veri kümeleri, beş farklı türde sınıflandırma algoritmasına tabi tutulmuştur. Elde edilen sonuçlar, boyut indirgeme için soyut öznitelik çıkarım yöntemi kullanıldığında diğer yöntemlere nazaran çok daha yüksek sınıflandırma başarımı elde edildiğini göstermektedir.

ABSTRACT

Classification of datasets that contain samples with numerous features is known as a costly process in time and space. In order to overcome this problem, dimensionality reduction techniques like feature selection and feature extraction are proposed in literature. In this paper, we compare the impacts of abstract feature extraction method and other popular techniques that use class labels for dimensionality reduction on classification performances. For evaluation, we utilize two standard text datasets having high dimensional samples. We compare the impacts of selected methods on performance by applying them on selected datasets and testing on five different classifiers with different design approaches. Results show that using abstract feature extraction method for dimensionality reduction produces much better classification performance, when compared with other selected methods.

1. GİRİŞ

Benzer özniteliklere sahip örnekleri önceden belirli olan kategorilere atama işlemine sınıflandırma adı verilir. Bilgiye erişimi kolaylaştırmak için; kütüphanecilik, biyoloji, tıbbi bilimler, yapay zeka gibi çok çeşitli disiplinlerde sınıflandırma çözümleri geliştirilmiştir. Yapay zeka alanında sınıflandırma çözümleri ses tanıma, haraketli ve sabit görüntü tanıma, belge sınıflama ve ayırt etme, kişisel özellikleri belirleme gibi değişik amaçlara ilişkindir. Ancak bu alandaki verilerin yapısı, algoritmaların etkin çalışmasını engelleyecek sayı ve boyutta özniteliklere sahiptir. Bu bildiride, yapay zeka alanındaki sınıflandırma problemlerine çözüm üretmeyi hedefleyen algoritmaların zaman ve uzay problemlerini azaltmak üzere ortaya atılmış olan, sınıf bilgisini kullanan boyut indirgeme yöntemlerinin başarıma olan etkisi, örnek olarak seçilen metin sınıflandırma problemi üzerinde karşılaştırılmaktadır. Metin sınıflandırma, belgeleri önceden tanımlı ve belirli sayıdaki kategoriye atama işlemi olarak tanımlanır [1].

Bu bildirinin ikinci bölümünde metin tipindeki verinin genel yapısı ve metin sınıflandırma için boyut indirgeme işleminden bahsedilecektir. Üçüncü bölümde seçilen boyut indirgeme yöntemleri kısaca tanıtılacaktır. Dördüncü bölümde yöntemleri karşılaştırmak üzere seçilen veri kümeleri tanıtılarak, deney kurulumu ve elde edilen sonuçlar verilecektir. Sonuç bölümünde ise elde edilen sonuçlar değerlendirilecektir.

2. METİN SINIFLANDIRMA İÇİN BOYUT İNDİRGEME

Sınıflandırılacak bir örnek, öznitelik vektörü olarak bilinen bir öznitelik kümesi şeklinde temsil edilir. Görüntü işlemede piksel bilgileri, biyoinformatikte DNA ya da protein dizileri şeklinde olabilen öznitelik vektörleri, metin işleme alanında vektör uzayı modeli olarak da bilinen terimlerin frekansları şeklinde ifade edilir [2]. Vektör uzayı modelinde her bir belge ya da metin bir örneğe karşılık gelmektedir. Belgeler, her bir boyutu belgedeki kelimelerden oluşan vektörler şeklinde ifade edilirler. Bir belgenin içerdiği kelimeler (terimler) onun öznitelik sayısını gösterir. Metin tipindeki veriler, yapısında çok sayıda öznitelik barındırmaktadır. Bundan dolayı metin sınıflandırma uzay ve zaman karmaşıklığı yüksek olan bir problemdir [3]. Bu problemin çözümü içinse kategorize edilecek verinin öznitelik boyutlarını indirgeme yöntemleri ortaya atılmıştır. Boyut indirgemede ilk yaklaşım öznitelik seçimidir ve başarımı en az düşürecek ve efektif çalışmayı arttıracak şekilde, özniteliklerin sayısının azaltılması hedeflenir. İkinci yaklaşım olan öznitelik çıkarımında ise amaç az sayıda yeni öznitelikle verinin yeniden tanımlanmasıdır.

2.1. Öznitelik Seçim Yöntemleri

Öznitelik seçim yöntemleri ile belgeleri diğerlerinden daha iyi tanımlayan terimler seçilmeye çalışılır. Bunun için çeşitli deneyler yaparak diğerlerinden daha iyi sonuç veren terim alt

978-1-4673-0056-8/12/$26.00 ©2012 IEEE

kümesini arayan yöntemler olduğu gibi, çeşitli değerlendirme ve dizme yöntemleriyle terimleri sıralayıp belirli bir eşik değerinin üzerinde değer alan terimleri seçen yöntemler de mevcuttur. Metin işleme uygulamalarında boyut indirgeme için genellikle öznitelik seçim yöntemleri tercih edilmektedir.
En bilinen ve en çok kullanılan öznitelik seçimi algoritmaları arasında belge frekansı (document frequency), chi istatistiği (chi statistic), bilgi kazanımı (information gain), terim dayanıklılığı (term strength) ve karşılıklı bilgi (mutual information) yer alır [4]. Sayılan yöntemlerin hepsi filtre yöntemlerdir ve öznitelikleri birbiriyle benzeşen entropi temelli değer hesabına göre süzerler. Ayrıca, yine popüler metotlar olan chi kare (chisquare) ve korelâsyon katsayısı (correlation coefficient) metotlarının belge frekansından daha iyi sonuçlar verdiği de ispatlanmıştır [5]. Tümü doğrusal olan bu yöntemlerin yanında, doğrusal olmayan RELIEF ve doğrusal olmayan çekirdek çarpımsal artımları (nonlinear kernel multiplicative updates) yöntemleri verilebilir [6]. Bu yöntemler arasından öznitelikleri seçerken sınıf bilgisini kullananlar chi kare ve RELIEF yöntemleridir. Bu çalışmada sınıf bilgisini kullanan yöntemler karşılaştırıldığı için sayılan iki öznitelik seçim yöntemi karşılaştırılmak üzere seçilmiştir.

2.2. Öznitelik Çıkarım Yöntemleri

Öznitelik çıkarım yöntemleri, belgeleri terimlerin bileşkesini alarak daha düşük boyutlu yeni bir uzayda kaynaştırılmış yeni özniteliklerle ifade eder. Bu sayede veri, sayıca daha az ve orijinallerinden bağımsız özniteliklerle ifade edilmiş olur.
Bilinen ve en yaygın olarak kullanılan boyut indirgeme yöntemi saklı anlamsal çözümlemedir (latent semantic analysis, LSA) [7]. Bunun dışında temel bileşen çözümlemesi (principal component analyis, PCA), çok boyutlu ölçekleme (multidimensional scaling, MDS), öğrenen yöney nicemleme (learning vector quantization, LVQ), doğrusal ayırtaç çözümlemesi (linear discriminant analysis, LDA), etken çözümlemesi (factor analysis, FA) gibi pek çok genel amaçlı öznitelik çıkarım yöntemi de metin işleme alanında uygulanmıştır [8]. Metin işleme alanı için geliştirilmiş soyut öznitelik çıkarım yöntemi (abstract feature extraction, AFE), günümüzde gerçekleştirilen güncel çalışmalardandır [9]. Sayılanlar arasında öznitelik çıkarırken sınıf bilgisini kullanan yöntemler doğrusal ayırtaç çözümlemesi (LDA) ve soyut öznitelik çıkarımı olduğundan, bu çalışmada karşılaştırmaya dahil edilmişlerdir.

3. SEÇİLEN BOYUT İNDİRGEME YÖNTEMLERİ

Boyut indirgeme için sınıf bilgisinden yararlanan öznitelik seçim yöntemleri olarak chi kare ve RELIEF, öznitelik çıkarım yöntemleri olarak da doğrusal ayırtaç çözümlemesi ve soyut öznitelik çıkarımı seçilmiştir. Amacımız, bu yöntemlerin metin sınıflandırma performansı üzerindeki etkilerini karşılaştırmaktır.

3.1. Chi Kare Öznitelik Seçim Yöntemi

Bu yöntemde özniteliklerin sınıflara göre chi kare istatistikleri hesaplanır. Bir sınıfta yer alan bir öznitelik için hesaplanan chi kare skoru, o terim ile o sınıf arasındaki bağımlılığı ölçmektedir. Eğer öznitelik sınıftan bağımsızsa, skoru sıfır olur. Yüksek bir chi kare skoruna sahip olan öznitelik, sınıf için daha tanımlayıcıdır.

N adet belgeden oluşan bir veri kümesinde, c_i sınıfında yer alan t terimi için χ^2 chi kare skoru,

$$\chi^2(t,c_i) = \frac{N[P(t,c_i)P(\bar{t},\bar{c}_i) - P(t,\bar{c}_i)P(\bar{t},c_i)]^2}{P(t)P(\bar{t})P(c_i)P(\bar{c}_i)} \quad (1)$$

ile hesaplanır [10].Burada t ve $\bar{t}$, terimin belgede var olup olmamasını, c_i ve $\bar{c}_i$ ise belgenin c sınıfında olup olmaması durumunu göstermektedir.
Veri kümesindeki tüm terimler için hesaplanan chi kare istatistik değerleri en yüksek olan k adet terim seçilerek diğerleri atıldığında, veri kümesi seçilen k adet terim ile ifade edilmiş olur.

3.2. RELIEF Öznitelik Seçim Yöntemi

RELIEF yöntemi, özniteliklerin değerini aralarında bulunan ya da bulunmayan bağımlılıkları ortaya çıkarmaya çalışarak bulmayı hedefler. Bunu yapmak için, özelliğin bulunduğu örneğin ait olduğu ve olmadığı sınıflarda yer alan en yakın örnekleri ağırlıklandırarak karşılaştırır. Orijinalinde iki sınıflı problemler için geliştirilen algoritma, ReliefF şeklinde çok sınıflı problemlere uyarlanmıştır [11]. Bir R örneğindeki A özelliği için $W[A]$ ağırlığı, ait olduğu sınıftaki en yakın örnek H ve ait olmadığı C adet sınıftaki en yakın örnekler $M(C)$ olduğunda,

$$W[A] = W[A] - diff(A,R,H)/m + \sum_{C \neq class(R)} [P(c) \times diff(A,R.M(C))]/m \quad (2)$$

ile hesaplanır. Burada m, normalizasyon katsayısı, *diff* ise iki örnek arasındaki fark fonksiyonudur.

3.3. LDA Öznitelik Çıkarım Yöntemi

LDA sınıflar arası ayrımı belirleyecek ayırtaçlar oluşturmak için orijinal özniteliklerin doğrusal bir bileşkesini oluşturur. Amaç sınıflar arası varyans ile sınıf içi varyans arasındaki oranı en yükseğe çıkarmaktır. w yönündeki sınıf ayrımı

$$S = \frac{w^T \sum_c (\mu_c - \mu)(\mu_c - \mu)^T w}{w^T \sum_c \sum_{i \in c} (x_i - \mu_c)(x_i - \mu_c)^T w} \quad (3)$$

ile hesaplanabilir [12]. Bu denklemlerde μ_c, c sınıfının ortalaması, μ ise tüm sınıfların ortalamalarının ortalamasıdır.
LDA ile elde edilen dönüşüm 3.19'de verilen sınıf ayrımını maksimuma çıkartır. Eğer w Σ_w^{-1} Σ_B matrisinin bir özvektörü ise, sınıf ayırtaçları özdeğerlere eşit olur ve doğrusal dönüşüm matrisi U

$$\begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_k \end{bmatrix} = \begin{bmatrix} u_1^T \\ u_2^T \\ \dots \\ u_k^T \end{bmatrix} (x - \mu) = U^T (x - \mu) \quad (4)$$

ile hesaplanabilir. U matrisinin kolonları Σ_w^{-1} Σ_B matrisinin özvektörlerinden oluşur.

$$\Sigma_B u_k = \lambda_k \Sigma_w u_k \quad (5)$$

denkleminin çözülmesi ile elde edilen özvektörler, öznitelikler arası değişimleri ifade eden bir vektör alt uzayı oluşturduğu için boyut indirgeme amacıyla kullanılır.

Doğrusal ayırtaçlar *c-1* adet olduğu için, sonuçta *c-1* boyuta indirgeme sağlanmış olur.

3.4. Soyut Öznitelik Çıkarım Yöntemi

Soyut öznitelik çıkarım yöntemi özvektörleri ya da özdeğerleri kullanmadığı gibi tekil değer ayrıştırımı da yapmaz, bu yönüyle literatürde yer alan diğer öznitelik çıkarımı yöntemlerinden farklıdır. Bu yöntemde terimlerin ağırlıkları ve sınıflar üzerindeki olasılıksal dağılımları göz önüne alınmaktadır. Terim olasılıklarının sınıflara olan izdüşümünü alıp bu olasılıkları toplayarak, her terimin sınıfları ne kadar etkilediği bulunmaktadır.

I adet terim, *J* adet belge ve *K* adet sınıf varken $n_{i,j}$ t_i teriminin d_j belgesinde kaç kere geçtiği, J_i de veri kümesinde t_i terimine sahip olan belge sayısı ise, soyut öznitelik çıkarımı

$$nc_{i,k} = \sum_{j} n_{i,j} \quad , \quad d_j \in c_k \tag{6}$$

$$w_{i,k} = \log(nc_{i,k} + 1) \times \log(\frac{J}{J_i}) \tag{7}$$

$$Y_{j,k} = \sum_{i} w_{i,k} \quad , \quad t_i \in d_j \tag{8}$$

$$AF_{j,k} = \frac{Y_{j,k}}{\sum_{k} Y_{j,k}} \tag{9}$$

ile hesaplanır. Sonuçta, *I* adet terim *K* boyutlu bir uzaya yansıtılır.

4. DENEY KURULUMU VE KARŞILAŞTIRMALI SONUÇLAR

Boyut indirgeme için sınıf bilgisini kullanan yöntemlerin sınıflandırma performansına olan etkilerini test etmek üzere, metin sınıflandırma işlemlerinde standart test veri kümeleri olarak bilinen ve kullanılan iki veri kümesi üzerinde sınıflandırma testleri gerçekleştirilmiştir.

4.1. Veri Kümeleri

Testlerde kullanılan ilk veri kümesi homojen dağılımlı 20-Newsgroups veri kümesidir. İkinci veri kümesi ise Reuters-21578 veri kümesinin heterojen yapısı ile bilinen ModApte-10 versiyonudur. Her iki veri kümesi de İngilizce metinlerden oluşmaktadır.

ModApte-10 veri kümesi eğitim ve test örnekleri olmak üzere hazır olarak ikiye bölünmüş durumdadır. Veri kümesinde 9603 eğitim, 3299 test örneği yer almaktadır. ModApte-10 aşırı derecede heterojen bir yapıya sahiptir [13]. Bir sınıfta çok az eğitim ve test örneği varken, başka bir sınıfta çok sayıda örnek bulunabilmektedir.

20 Newsgroups veri kümesi, metin sınıflandırma ve kümeleme işlemleri için sıklıkla kullanılan bir veri kümesidir [14]. 20 değişik başlık altında toplanmış yaklaşık 20.000 adet haber grubu postası koleksiyonundan oluşmaktadır. Genel olarak veri kümesindeki başlıkların 6 grupta toplandığı görülmektedir. Çalışmada tüm veri kümesi yerine, her sınıfta yüzer adet örnek barındıran "20 Newsgroups Mini" [15] versiyonu kullanılmıştır.

Veri kümelerini, boyutlarını seçilen yöntemlerle indirgeyerek sınıflandırma performans karşılaştırma testlerine hazırlamak üzere, Porter'ın [16] kök bulma yöntemi ile her iki veri kümesindeki belgelerde yer alan kelimelerin kökleri bulunmuştur. İngilizce sık kullanılan kelimeler temizlenmiş, sayılar ve noktalama işaretleri kaldırılmıştır. Bu aşamalardan geçerek ön işlemeleri gerçekleştirilen veri kümelerinin, ikinci bölümde detayları verilen yöntemlerle boyutları indirgenmiştir. Elde edilen indirgenmiş veri kümelerinin yeni boyutları Tablo 1'de verilmiştir.

Tablo 1: İndirgenmiş veri kümelerinin boyutları

Yöntem	20-Newsgroups	ModApte-10
Chi kare	326	2019
Relief	5590	4489
LDA	19	9
AFE	20	10

4.2. Deney Kurulumu

Seçilen yöntemlerin sınıflandırma performansına etkisini test etmek üzere istatistik temelli, karar ağacı, kural tabanlı, örnek temelli ve çekirdek tabanlı sınıflandırıcı türlerinin en bilinen ve kullanılan algoritmalarına yer verilmiştir. Tüm algoritmaların standart parametreleri kullanılmıştır.

İstatistikî sınıflandırıcı olarak sınıflandırıcısı Bayes teoremini güçlü bağımsız varsayımlara uygulamaya dayalı olan Naive Bayes kullanılmıştır [17].

Quinlan'ın [18] C4.5 algoritmasının uygulaması olan J48 ağacı, karar ağacı sınıflandırıcısına örnek olarak seçilmiştir.

Örnek temelli sınıflandırıcılardan 1 en yakın komşu (1-NN) algoritması seçilmiştir.

Kural tabanlı sınıflandırıcı olarak RIPPER algoritması kullanılmıştır [19].

Verinin seyrekliğine dayanıklı çekirdek tabanlı sınıflandırıcı olarak doğrusal çekirdeğe sahip destek vektör makineleri (Support Vector Machines, SVM) [20] seçilmiştir.

Yöntemleri karşılaştırmak üzere yapılan testlerde doğrulama yöntemi olarak 20-Newsgroups veri kümesinde 10 kere çapraz doğrulama (10-fold cross validation) kullanılmıştır. 10 kere çapraz doğrulamanın en önemli avantajı, veri kümesindeki tüm örneklerin eğitim ve test aşamalarında kullanılmasıdır. Bu sayede bazı örneklerin eğitim sürecinde yapmış olabilecekleri pozitif veya negatif etkiler bertaraf edilmiş olur. ModApte-10 veri kümesi ise standart olarak eğitim ve test örnekleri şeklinde ayrık durumdadır. Bu veri kümesinde ayrık eğitim ve test kümeleri olduğu halleriyle kullanılmıştır.

4.3. Test Sonuçları

Sınıf bilgisini kullanan boyut indirgeme yöntemlerinin sınıflandırma performansına olan etkisini karşılaştırmak üzere seçilen veri kümeleri kullanılarak yapılan testlerde performans, duyarlılık (precision, true positive/(true positive+false positive)) ve anma (recall, true positive/(true positive+false negative)) ölçeklerinin harmonik ortalaması olan ve

$$F_1 = \frac{2 \times precision \times recall}{(precision + recall)} \tag{10}$$

ile hesaplanan F_1 ölceği (F_1 measure) ile ölçülmüştür.

Yapılan testlerde elde edilen F_1 ölçeği sonuçları 20-Newsgroups veri kümesi için Tablo 2'de, ModApte-10 veri kümesi için Tablo 3'te verilmiştir. Tablolarda her satırın en

iyi değeri koyu, her sütunun en iyi değeri ise eğik yazılarak belirtilmiştir.

Tablo 2: 20-Newsgroups testlerinin F_1 ölçeği sonuçları

Sın.Alg.	Chi Kare	Relief	LDA	AFE
N.Bayes	0.597	0.594	0.289	**0.897**
C4.5	0.500	0.468	0.363	**0.869**
1-NN	0.489	0.280	0.311	**0.922**
RIPPER	0.467	0.434	0.343	**0.877**
SVM	*0.631*	*0.697*	*0.409*	***0.930***
Ortalama	0.537	0.495	0.343	**0.899**

Tablo 3: ModApte-10 testlerinin F_1 ölçeği sonuçları

Sın.Alg.	Chi Kare	Relief	LDA	AFE
N.Bayes	0.803	0.739	0.732	**0.911**
C4.5	0.881	0.882	0.805	**0.948**
1-NN	0.659	0.784	0.775	***0.956***
RIPPER	0.863	0.867	0.767	**0.949**
SVM	*0.914*	***0.919***	*0.808*	0.882
Ortalama	0.824	0.838	0.777	**0.929**

20-Newsgroups veri kümesindeki testlerde tüm algoritmalarda en yüksek başarım, veri kümesi AFE ile indirgendiğinde elde edilmiştir. ModApte-10 veri kümesindeki testlerde, RELIEF ile indirgenip SVM ile sınıflandırma hariç tüm algoritmalarda yine AFE ile boyut indirgeme en yüksek başarımları sağlamıştır. Sınıflandırma algoritmalarının genel başarımlarına bakıldığında ise; testlerin biri hariç (AFE ile indirgenmiş Modapte-10 veri kümesi) hepsinde SVM sınıflandırma algoritması en iyi başarımı sağlamıştır. Veri kümelerinde elde en yüksek performanslar AFE ile indirgendiğinde elde edilmiş; 20-Newsgroups veri kümesinde SVM, ModApte-10 veri kümesinde ise 1 en yakın komşu algoritmalarında sağlanmıştır.

5. SONUÇLAR

Metin sınıflandırmada performansı etkileyen ve süreci zorlaştıran en önemli engel verinin yüksek boyutlu olmasıdır. Çok sayıda terimle ifade edilen belgeler üzerinde sınıflandırma için gereken işlem gücü ve kaynak miktarı çoğu zaman bu işlemlerin yapılmasını güçleştirmektedir. Yüksek boyutluluğun getirdiği sorun için uygulanan çözüm, verinin boyutlarının indirgenmesidir. Bu çalışmada boyut indirgeme için sınıf bilgisini kullanan yöntemler kısaca tanıtılarak örnek veri kümeleri üzerinde uygulanmış ve sınıflandırma performansına olan etkileri karşılaştırılmıştır. Test sonuçları karşılaştırılarak incelendiğinde, boyut indirgeme için soyut öznitelik çıkarımı yönteminin biri hariç tüm testlerde en yüksek başarımı sağladığı görülmüştür. Boyutları indirgenmiş veri kümelerindeki testlerin biri hariç hepsinde en başarılı sınıflandırıcı SVM olmuştur. Bu konuda yapılacak gelecek çalışma, yöntemleri metin işleme dışında görüntü işleme, ses işleme gibi alanlarda da karşılaştırmak olacaktır.

6. KAYNAKÇA

[1] Joachims, T., "*A Probabilistic Analysis of the Rocchio Algorithm with Tfidf for Text Categorization*" in Proc.14th Int. Conf. on Machine Learning, Nashville, 1997, pp. 143-151.

[2] Efron, M. "Query Expansion and dimensionality reduction: Notions of optimality in Rocchio Relevance Feedback and Latent Semantic Analysis," *Information Processing & Management*, vol. 44(1):163-180, 2008.

[3] Bunte, K., et al., "Adaptive Local Dissimilarity Measures for discriminative dimension reduction of labeled data," *Neurocomputing*, vol. 73(7-9):1074-1092, 2010.

[4] Zhu, J., Wang, H., ve Zhang, X., "Discrimination-Based Feature Selection for Multinomial Naïve Bayes Text Classification", *LNAI*, vol.4285:149-156, 2006.

[5] Jensen, R., Shen, Q., *Computational Intelligence and Feature Selection Rough and Fuzzy Approaches,* IEEE-Wiley, New Jersey, 2008.

[6] Guyon, I., et al., "Multivariate Non-Linear Feature Selection with Kernel Methods", *Studies in Fuzziness and Soft Computing*, 164:313-326, 2005.

[7] Landauer, T. K., Dumais, S. T., "A Solution to Pluto's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge", *Psychological Review*, 104(2): 211–240, 1997.

[8] Fodor, I. (Haziran 2002). *A Survey of Dimension Reduction Techniques* [online]. https://computation.llnl.gov/casc/sapphire/pubs/148494.pdf

[9] Biricik, G., "Metin Sınıflama için Yeni Bir Özellik Çıkarım Yöntemi", Doktora tezi, Yıldız Teknik Üniversitesi, İstanbul, 2011.

[10] Zheng, Z., Wu, X. ve Srihari, R., "Feature Selection for Text Categorization on Imbalanced Data", *ACM SIGKDD Exploraitons Newsletter*, 6(1):80-89, 2004.

[11] Kononenko, I., "*Estimating Attributes: Analysis and Extensions of RELIEF*" European Conference on Machine Learning, Catania, Italy, 1994, pp. 171-182.

[12] Martinez, A. M. ve Kak, A.C., "PCA versus LDA", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(2):228-233, 2001.

[13] *ModApte Split of Reuters-21578 Dataset* [online]. http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-text-classification-1.html.

[14] *20 Newsgroups veri kümesi* [online]. http://people.csail.mit.edu/jrennie/20Newsgroups

[15] *20 Newsgroups özet veri kümesi* [online]. http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html.

[16] Porter, M.F., "An Algorithm for Suffix Stripping", *Program*, 14(3):130-137, 1980.

[17] McCallum, A., Nigam, K., "*A Comparison of Event Models for Naive Bayes Text Classification*". Proceedings of AAAI/ICML-98 Workshop on Learning for Text Categorization, Madison, Wisconsin, 1998, pp.41-48.

[18] Quinlan, J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, California, 1993.

[19] Cohen, W.W., "*Fast Effective Rule Induction*", 12th Int. Conf. on Machine Learning, Tahoe City, California, 1995, pp.115-123.

[20] Cortes, C., Vapnik, V., "Support-Vector Networks", *Machine Learning*, 20(3):273-297, 1995.