SOYUT ÖZELLİK ÇIKARIMI İLE YAZAR TANIMA
AUTHOR RECOGNITION BY ABSTRACT FEATURE EXTRACTION

Murat Yasdi
Yıldız Teknik Üniversitesi
Bilgisayar Mühendisliği Bölümü
murat.yasdi@gmail.com

Banu Diri
Yıldız Teknik Üniversitesi
Bilgisayar Mühendisliği Bölümü
banu@ce.yildiz.edu.tr

ÖZETÇE

Bu çalışmanın amacı Soyut Özellik Çıkarım yönteminin çok boyutlu özellik vektörlerine sahip olan çalışmalarda başarılı sonuçlar verdiğini göstermektir. Uygulama alanı olarak yazar tanıma çalışması alınmış ve özellik vektörleri olarak da kelime kökleri ve 2 gram'lar seçilmiştir. Soyut Özellik Çıkarım yönteminin başarısı PCA, CFS, ki-kare gibi özellik çıkarım yöntemleri ile kıyaslanarak sınıflandırmadaki başarısı hem Türkçe hem de İngilizce veri setleri üzerinde gösterilmiştir.

ABSTRACT

The purpose of this study is to show the success of Abstract Feature Extraction Method in multi dimensional feature vectors studies. Author recognition study is taken as an application area and word root and 2 gram's are chosen as feature vectors. The sucess of the Abstract Feature Extraction method in classification is shown on both Turkish and English data sets by comparing with feature extraction methods such as PCA, CFS, chi-square.

1. GİRİŞ

Doküman sınıflandırma, bir dokümanın sahip olduğu özelliklere bakılarak önceden belirlenmiş belli sayıdaki kategorilerden hangisine dahil olacağını belirlemektir. Doküman sınıflandırmadaki problemlerden biri, kime ait olduğu bilinmeyen veya yazarının kimliğinden şüphelenilen dokümanların yazarının tahmin edilmesi, bir diğer problem de dokümanın türünün veya yazarının cinsiyetinin belirlenmesidir.

Doküman sınıflandırma üzerine yapılan ilk çalışmalar yetmişli yıllarda otomatik doküman indeksleme olarak karşımıza çıkmıştır. Belirli konular için özel sözlükler oluşturulmuş ve sözlük içerisindeki kelimeler de birer kategori olarak kabul edilerek dokümanlar sınıflandırılmıştır. Yazar tanıma üzerine yapılan çalışmalardan Mosteller [1] yazarlık özelliklerini çıkarmış ve Bayesian analizi ile tanıma yapmışlardır. Burrows [2] ise çalışmasında en fazla sıklıkta kullanılan kelimeleri, Brinegar [3] kelimelerin uzunluğunu, Morton [4] cümlelerin uzunluğunu, Brainerd [5] ortalama hece sayısını, Holmes [6] kullanılan kelime sayısını ve dokümanın uzunluğunu, Twedie [7] farklı kelime sayısının toplam kelime sayısına oranını kullanmıştır. Stamatatos [8] doğal dil işleme hazır paket programını kullanarak bir dizi stil belirleyici (style marker) elde etmiş ve bunlardan yararlanarak yazar tanıma yapmışlardır. Fürnkranz [9] 2-gram ve 3-gram özelliklerini, Tan [10] sadece 2-gram'ları kullanarak doküman sınıflandırmada performansı arttırmışlardır. Çatal [11], n-gram'ları kullanarak NECL adını verdikleri bir sistem geliştirmişlerdir. Diri [12] bir dokümanın yazarını ve türünü belirlemede kullanılmak üzere 22 adet stil belirleyicisi çıkarmış ve bunları kullanan bir sınıflandırma sistemi geliştirmişlerdir. Yine aynı yazarlar [13] 2 ve 3-gram'ları kullanarak dokümanın yazarını, türünü ve yazarının cinsiyetini belirme üzerine çalışmışlardır. Ganiz [14] ise doküman sınıflandırmada yazıya ön işlem uygulamanın performansı arttırdığını göstermiştir.

Bu çalışmada yazar belirlemede kullanılan özellikler olarak kelime kökleri ve 2-gram lar özellik olarak alınmıştır. Bu özelliklerden de en ayırt edici olanları Soyut Özellik Çıkarımı yöntemi ile belirlenerek tanıma işleminde kullanılmıştır. Çalışmanın ilerleyen bölümlerinde yazar tanıma sisteminin adımları, özelliklerin çıkarılması, ayırtedici özelliklerin seçilmesi, veri setinin tanıtılması ve deneysel sonuçlar olarak sunulacaktır.

2. YAZAR TANIMA SİSTEMİ

Geliştirilen yazar tanıma sisteminin dilden bağımsız olarak çalışması düşünülmüş ve deneysel sonuçlar Türkçe ve İngilizce için yapılmıştır. Tanımada kullanılan özellik olarak kelime kökleri ve karakter n-gram'lar (n=2) seçilmiştir.

Kelime Kökleri

Çalışılan dillerden biri olan Türkçe sondan eklemeli bir dil olduğu için *defterlerinin*, *defterlerin*, *defterleri* gibi üç farklı kelime yerine sadece bu kelimelerin kökü olan *defter* kelimesinin alınması seçilecek özelliklerin belirlenmesinde önemldir. Bu yüzden şekil-1'de de görüldüğü gibi Türkçe dili için Zemberek[1], İngilizce için de Porter Stemmer[2] kullanılmıştır. Kelime köklerinin belirlenmesinde doküman içerisinde yer alan etkisiz kelimelerin (stop words) çıkarılması veya dahil edilmesi kullanıcıya seçimlik olarak bırakılmıştır.

N – Gramlar

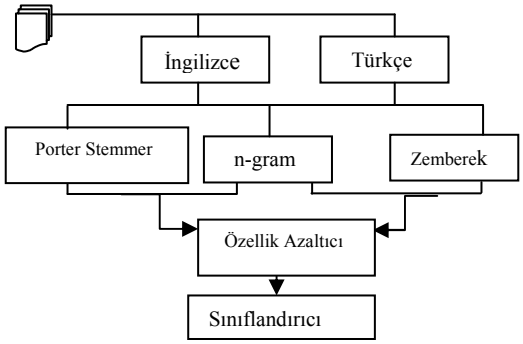
n-gram'ların sınıflandırma problemlerinde [13] başarılı sonuçlar vermesinin yanısıra özellik olarak seçilmesinin diğer bir sebebi yazar tanıma sisteminin dilden bağımsız olarak tasarlanacak olmasıdır. Ayrıca, bu çalışma

[1] http://code.google.com/p/zemberek/
[2] http://tartarus.org/~martin/PorterStemmer/

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

içerisinde yer alan Soyut Özellik Çıkarım yönteminin özellik azaltmada ki başarısını göstermek amacıyla tercih edilmiştir.

Özellik Frekansları ve Trie Yapısı



Şekil 1: Sistemin genel yapısı

Kelime gövdelerinin ve n-gram'ların dokümanlarda geçme sıklıklarını ve kaç farklı dokümanda geçtiği bilgisini bulmak için Trie ağaç yapısı kullanılmıştır. Trie ağacı, düğümlerinde kelimelerin harflerini sıralı bir şekilde tutar ve kelimelerin son harfini tutan düğümlerinde de o kelimenin geçiş frekansını saklar. Ağaca yeni bir özellik eklerken ağaçta harf harf gezilerek kelimenin son harfine kadar ilerlenir, eğer kelimenin son harfi bir düğüme karşılık geliyorsa, bu düğümün bilgisi güncellenir, gelmiyorsa ağaca yeni bir düğüm eklenir ve başlangıç değeri verilir. Ağaca yeni bir özellik eklemek ve aramak bu yöntem ile çok hızlı bir şekilde gerçekleştirilir.

2.1. Soyut Özellik Çıkarım Yöntemi

Soyut Özellik Çıkarım (AFE-Abstract Feature Extraction) yöntemi ne öz değer/öz vektörleri kullanır ne de tekil değer ayrıştırımı yapar. Bu yöntem terimlerin ağırlıklarını ve sınıflar üzerindeki olasılıksal dağılımlarını göz önüne alarak çalışır. Terim olasılıklarının sınıflara olan izdüşümü alınıp, bu olasılıklar toplanarak her terimin sınıfları ne kadar etkilediği hesaplanır [15, 16].

I adet terim, J adet doküman ve K adet sınıf olduğu kabul edilirse $n_{i,j}$, t_i teriminin d_j dokümanında kaç kere geçtiğini göstersin. N_i de, t_i terimine sahip olan doküman sayısı olsun. Bir t_i teriminin c_k sınıfında kaç kere geçtiği denklem (1) ile hesaplanabilir. Bir d_j dokümanında yer alan t_i teriminin c_k sınıfını ne kadar etkilediği ise denklem (2) ile bulunabilir.

$$nc_{i,k} = \sum_j n_{i,j}\,, \qquad d_j \in c_k \tag{1}$$

$$w_{i,k} = log\left(nc_{i,k} + 1\right) \times log\left(\frac{N}{N_i}\right) \tag{2}$$

Bir dokümandaki tüm terimlerin ck sınıfına olan toplam etkisi ise denklem (3) ile elde edilir. Sonuçta, I adet terim sınıf sayısı kadar boyutlu bir hiper düzleme yansıtılır. Bu işlem tüm dokümanlar için uygulandığında J satır (her belge için bir satır) ve K sütundan oluşan (çıkarılan özelliklerin sayısı sınıf sayısına eşittir) indirgenmiş sonuç matrisi elde edilir. Son adımda da denklem (4) ile yeni terimler normalize edilir.

$$Y_k = \sum_i w_{i,k}\,, \qquad w_i \in d_j \tag{3}$$

$$YeniTerim_k = \frac{Y_k}{\sum_k Y_k} \tag{4}$$

Soyut Özellik Çıkarımı yöntemi aynı zamanda bir sınıflandırıcı olarak da kullanılabilir.

2.2. Özellik Seçme ve Sınıflandırma Yöntemleri

Temel Bileşen Analizi (PCA-Principal Component Analysis) veri setinin varyansını en iyi yansıtan ve bilginin önemli yönlerini gösteren özellikler olan alt düzey temel bileşenlerin karakteristiğini koruyarak özellik çıkarımı yapar. Bu da özellik kümesinin yeni bir hiper uzaya iz düşümünün öz değerleri ve öz vektörlerinin hesaplanması ile gerçekleşir.

Korelasyon Tabanlı Özellik Seçici (Correlation-based Feature Selection-CFS) Diğer özelliklerle düşük korelasyonlu, sınıf değişkeni ile yüksek korelasyonlu olan özellikleri seçer.

Ki-Kare Yöntemi Özelliklerin sınıflara göre ki-kare istatistiklerine bakılarak her özellik teker teker değerlendirilir. Bir sınıfta yer alan bir özellik için hesaplanan Ki-kare değeri, o terim ile o sınıf arasındaki bağımlılığı ölçmektedir. Eğer özellik sınıftan bağımsız ise değeri sıfır olur. Yüksek Ki-kare değerine sahip olan özellik, sınıf için daha tanımlayıcıdır.

Naïve Bayes (NB) Klasik Naïve Bayes algoritması kelimelerin ve sınıfların birleşik olasılıkları ile bir dokümanın sınıfının belirlenmesinde kullanılır. Burada ise özellikler kelimelerin frekansları değildir ve sürekli dağılımlara sahip olduklarından klasik Naïve Bayes yerine George'un [17] çalışmasında önerilen Naïve Bayes versiyonu kullanılmıştır.

Destek Vektör Makinesi (DVM) Sınıfları birbirinden ayıran marjini en büyük, doğrusal bir ayırt edici fonksiyon bulunmasını amaçlar. Doğrusal olarak ayrılamayan örnekler için örnekler, doğrusal olarak ayrılabildikleri daha yüksek boyutlu başka bir uzaya taşınır ve sınıflandırma o uzayda yapılır.

Rasgele Orman (RO) Tek bir karar ağacı üretmek yerine her biri farklı eğitim kümeleriyle eğitilen çok sayıda, çok değişkenli ağaçların kararlarının birleştirilmesini kullanır. Farklı eğitim kümeleri önyükleme ve rasgele özellik seçimi ile orijinal eğitim setinden oluşturulur.

K-En Yakın Komşuluk (KEK) Sınıfı bulunacak olan verinin özellik vektörü değerine Euclidean mesafesi en yakın olan *k* adet özellik vektörlerinin bulunması prensibine dayanır. Bulunan bu *k* adet vektör en fazla hangi sınıfa ait ise, o sınıf etiketi sınıflandırılacak olan verinin sınıfı olarak belirlenir. Bölüm 2.2'de bahsedilen yöntemlerin hepsi Weka[1]'da varsayılan parametre

[1] www.cs.waikato.ac.nz/ml/weka/

değerleri ile çalıştırılmıştır. Sınıflandırma başarım değerleri de 10'lu çapraz geçerleme ile alınmıştır.

2.3. Veri Seti Özellikleri

Çalışmada Türkçe için kullanılan veri setleri günlük gazetelerin köşe yazılarından toplanmıştır. Üç farklı veri seti oluşturulmuştur. $VS_{tür}$ içerisinde sağlık, siyaset, spor ve magazin olmak üzere 4 türden oluşmaktadır. Her tür içerisinden de 5 yazar ve her yazarında 10 adet yazısı bulunmaktadır. VS_{cins} ikinci veri seti olup, içerisinde 10'u kadın, 10 tanesi de erkek olan 20 yazardan oluşmaktadır. Yine her yazarın 10 yazısı alınmıştır. Son veri seti VS_{yazar} olup, üçü siyaset, üçü magazin, ikisi spor ve ikisi sağlık olmak üzere farklı türlerde yazan 10 tane yazarın 10 yazısından oluşmakta olup, toplamda 100 yazıyı içerisinde bulundurmaktadır.

Soyut Özellik Çıkarım yönteminin dilden bağımsız da başarılı sonuçlar verdiğini göstermek amacıyla İngilizce için de Kemik[1] grubu tarafından hazırlanmış gazete yazarlarından 10 tanesi seçilerek VS_{ing} veri seti oluşturulmuştur. Bu yazarların cinsiyet ve yazı türü belirtilmemiş olduğundan sadece genel kategorisinde başarısı değerlendirilecektir.

3. DENEYSEL SONUÇLAR

Deneysel sonuçlar alınırken her veri seti için hem kelime köklerinden hem de 2 gramlar'dan oluşan 2 ayrı özellik vektörü elde edilmiştir. Toplamda sekiz farklı özellik vektörü herhangibir özellik azaltıcıya sokulmadan sınıflandırmadaki başarıları 10-katlı çapraz geçerleme ile alınmıştır. Deneysel sonuçların ikinci bölümünde ise elde edilen bu özellik vektörleri PCA, CfsSubsetEval (Cfs), Ki-kare ve Soyut Özellik Çıkarım (AFE) yöntemleri kullanılarak özellik sayıları azaltılarak yeniden sınıflandırma başarıları ölçülmüştür.

Tablo 1, 2 ve 3 sırasıyla VS_{cins}, $VS_{tür}$ ve VS_{yazar} göre alınmış başarıları göstermektedir.

Tablo 1: Cinsiyet göre yazar tanıma

Özellik Sayısı		NB	RO	DVM	KEK
Kelime kök	**4687**	69	70,5	78	51
	PCA				
	CFS(84)	90,5	82,5	89	76,5
	Ki-kare(296)	91	81	92	61
	AFE(2)	77	92	81,5	**94**
2-gram	**3163**	76	72,5	86	65
	PCA(162)	59	68,5	54,5	60
	CFS(74)	87	85	91	89,5
	Ki-kare(435)	88,5	78	**94,5**	79
	AFE(2)	68	81,5	72	85,5

VS_{cins} veri seti kullanılarak tüm yazılardan 4687 tane farklı kelime kökü ve 3163 tane de farklı 2-gram çıkarılmıştır. Kelime kökünden oluşan özellik vektöründen alınan en yüksek sınıflandırma başarısı DVM ile %78 iken, özellik vektörünün boyutu düşürüldüğünde sınıflandırma başarısında yükselme görülmektedir. (PCA yöntemi bazı veri setlerinde elimizdeki donanım imkanları dahilinde çalıştırılamamıştır). Cfs, Ki-kare ve AFE boyut azaltma yöntemleri kullanılarak sırasıyla özellik sayısı 84, 296 ve 2'ye düşürüldü. AFE, KEK ve RO sınıflandırıcılarında en yüksek performansı vermiştir. Kısaca, kelime köklerinden oluşan bu özellik vektörü cinsiyet bulma da AFE özellik azaltıcı ve k-En Yakın Komşuluk birlikte çalıştırıldığında %94'lük bir başarı elde edilmiştir. Aynı veri setinde, özellik vektörü olarak 2-gram'ları aldığımızda AFE'nin diğer özellik azaltıcı yöntemler arasında başarı sağlayamadığı görülmektedir. En iyi başarı %94,5 ile Ki-kare ve DVM ile alınmıştır. Yine de AFE'nin sadece 2 özellik ile aldığı başarı göz ardı edilemez.

Tablo 2: Türe göre yazar tanıma

Özellik Sayısı		NB	RO	DVM	KEK
Kelime kök	**4675**	86,5	75,5	92,5	32,5
	PCA				
	CFS(58)	91	88	89,5	82,5
	Ki-kare(431)	92,5	87	96,5	65,5
	AFE(4)	75,5	92,5	85	**97,5**
2-gram	**3289**	89	78,5	91,5	64,5
	PCA(143)	65	85,5	52	48
	CFS(60)	92,5	89,5	**95,5**	88,5
	Ki-kare(1144)	90	82	**95,5**	77
	AFE(4)	74	88,5	79	**95,5**

İçerisinde dört farklı tür bulunduran $VS_{tür}$ ile yapılan denemelerde kelime köklerinden oluşan özellikler 2-gram özelliklerden daha başarılı sonuç vermiştir. 4675 adet kelime kökü ile yapılan tür belirlemede DVM %92,5'lik bir başarı vermiştir. Özellik sayısını azalttığımızda Cfs, Ki-kare ve AFE sırasıyla 58, 431 ve 4 özellik sayısına inmiştir. En başarılı sonucu %97,5 ile AFE-KEK çifti vermiştir. 2-gram'lar ile yapılan tür belirleme de ise, Cfs ve Ki-kare, DVM ile AFE'de KEK ile %95,5'lik bir başarı vermiştir.

Tablo 3: Türden bağımsız (karışık) yazar tanıma

Özellik Sayısı		NB	RO	DVM	KEK
Kelime kök	**4729**	63	49	49	22
	PCA				
	CFS(33)	83	82	75	72
	Ki-kare(94)	79	79	84	71
	AFE(10)	81	91	91	**99**
2-gram	**2866**	84	64	94	48
	PCA(77)	71	75	41	48
	CFS(48)	**98**	91	92	91
	Ki-kare(626)	91	76	97	91
	AFE(10)	70	81	87	96

Olaya cinsiyet ve tür dışında sadece yazar tanıma olarak bakmak için oluşturulan VS_{yazar} veri seti içerisinde 10

[1] www.kemik.yildiz.edu.tr

farklı yazara ait yazılar bulunmaktadır. Burada da diğer veri setlerinde olduğu gibi AFE'nin özellik sayısı sınıf sayısı ile aynı olup, kelime köklerini kullanarak yapılan tanımada KEK yöntemi ile %99'luk başarı alınmıştır. 2-gram'larda ise Cfs özellik azaltıcı, NB ile %98'lik tanıma başarısı vermiştir. Bu özellik vektöründe AFE, KEK ile %96'lık bir başarı elde etmiştir.

Tablo 4'te ise VS_{ing} veri seti ile alınan sonuçlar görülmektedir. İngilizcede hem kelime kökleri hem de 2-gram'lar ile özellik azaltılmadan tanıma işlemi yapıldığında oldukça düşük sonuçlar alınmıştır. Her iki özellik vektöründe PCA kötü sonuçlar verirken, kelime köklerinde Cfs ve Ki-kare yöntemleri yaklaşık başarı vermiştir. AFE ile bütün sınıflandırıcılarda yüksek başarı alınmıştır. KEK ile alınan başarı %100'dür.

Tablo 4: VS_{ing} göre yazar tanıma

Özellik Sayısı		NB	RO	DVM	KEK
Kelime kök	**3728**	14	43	69	14
	PCA(90)	48	46	27	19
	CFS(25)	77	66	69	62
	Ki-kare(43)	78	66	69	62
	AFE(10)	93	93	94	**100**
2-gram	**1144**	53	42	92	16
	PCA(85)	59	55	24	13
	CFS(28)	81	87	87	78
	Ki-kare(62)	83	84	86	83
	AFE(10)	81	80	88	**97**

Özellik vektörü olarak 2-gram'lar kullanıldığında Cfs, Ki-kare ve AFE birbirlerine yakın sonuçlar vermesine rağmen, en yüksek başarı AFE-KEK ikilisinden %97 ile elde edilmiştir.

4. SONUÇLAR

Bu çalışmada esas amaç, Soyut Özellik Çıkarım yönteminin sınıflandırmadaki başarısını göstermektir. Uygulama alanı olarak yazar tanıma problemi alınarak yazarların cinsiyetini, yazıların türünü ve yazının sahibini tanımak amaçlı veri setleri hazırlanmış, kelime kökü ve 2-gram'lar çıkarılacak özellikler olarak belirlenerek özellik vektörleri oluşturulmuştur. Bu vektörlerin boyutları büyük olduğundan sınıflama başarımı düşük çıkmaktadır. Bu yüzden AFE'nin de aralarında olduğu farklı boyut azaltma yöntemleri kullanıldıktan sonra yapılan sınıflama işlemlerinden yüksek performanslar alınmıştır. AFE'nin, hem cinsiyet hem de tür ve yazar tanımada başarılı bir yöntem olduğu gösterilmiştir. Alınan başarının dilden bağımsız olduğunu göstermek içinde ayrıca ingilizce, bir veri seti oluşturularak denemeler yapılmıştır.

5. KAYNAKÇA

1. Mosteller, F., Wallace, D. L., *Applied Bayesian and Classical Inference: The Case of the Federalist Papers. Reading*, MA:Addison-Wesley, 1984
2. Burrows, J. F., "Not unless you ask nicely: the interpretative nexus between analysis and information", *Literary Linguist Computing, Vol. 7, 1992, p 91-109*
3. Stamatatos, E., Fakotakis, N., Kokkinakis, G., "Automatic Text Categorization in Terms of Genre and Author", *Computational Linguistics*, p 471-495, 2000
4. Morton, A. Q., "The Authorship of Greek Prose", *Journal of the Royal Statistical Society*, *Series A, 128:169-233, 1965*
5. Brainerd, B., *Weighting Evidence in Language and Literature: A Statistical Approach*, University of Toronto Press, 1974
6. Holmes, D. I, Authorship Attribution, *Computers and The Humanities, Vol.28, 1994, p 87-106*
7. Tweedie, F., Baayen, H., How Variable may a Constant be Measures of Lexical Richness in Perspective, *Computers and The Humanities, Vol. 32(5), 1998, p 323-352*
8. Stamatatos, E., Fakotakis, N., Kokkinakis, G., Computer-Based Authorship Attribution Without Lexical Measures, *Computers and The Humanities, Vol. 35, 2001, p 193-214*
9. Fürnkranz, J., *A Study using n-gram Features for Text Categorization*, Austrian Research Institute for Artifical Intelligence, 1998
10. Tan, C. M., Wang, Y. F., Lee, C. D., The Use of Bi-grams to Enhance, *Journal Information Processing and Management, Vol.30-4, 2002, p.529-546*
11. Çatal, Ç., Erbakırcı, K., Erenler, Y., "Computer-based Authorship Attribution for Turkish Documents", *Turkish Symposium on Artificial Intelligence and Neural Networks*, 2003
12. Diri, B., Amasyalı, M. F., "Automatic Author Detection for Turkish Texts", *Artificial Neural Networks and Neural Information Processing*, 138-141, 2003
13. Amasyalı, M. F., Diri, B., "Automatic Written Turkish Text Categorization in Terms of Author, Genre and Gender", *11th International Conference on Applications of Natural Language to Information Systems*, Austria, 2006
14. Torunoğlu, D., Çakırman, E., Ganiz, M. C., Akyokuş, S. and Gürbüz, M. Z., "Analysis of preprocessing methods on classification of Turkish texts", *Innovations in Intelligent Systems and Applications*, p 122-117, Istanbul, 2011
15. Biricik, G., Diri, B. and Sönmez, A: C., "A New Method For Attribute Extraction with Application on Text Classification", *5th International Conference on Soft Computing, Computing with Words*, ICSCCW, North Cyprus, Famagusta, 2009
16. Biricik, G., Diri, B. and Sönmez, A. C., "Impact of a New Attribute Extraction Algorithm on Web Page Classification", *5th Int. Conference on Data Mining*, DMIN'09, Las Vegas, USA, 2009
17. George, H., "Estimating Continuous Distributions in Bayesian Classifiers", *11th Conference on Uncertainty in Artificial Intelligence*, p 338-345, San Mateo, 1995