SÜREKLİ AKAN VERİLERDE AYKIRILIK TABANLI YENİ YORUM FİLTRELEME YAKLAŞIMI

NOVEL COMMENT FILTERING APPROACH BASED ON OUTLIER ON STREAMING DATA

Nuri Özalp, Güray Yılmaz

Bilgisayar Mühendisliği Bölümü
Hava Harp Okulu
nozalp@bte.tubitak.gov.tr, gyilmaz@hho.edu.tr

Uğur Ayan

Bilişim ve Bilgi Güvenliği İleri Teknolojiler
Araştırma Merkezi, TUBITAK
ugur.ayan@bte.tubitak.gov.tr

ÖZETÇE

Bu çalışma, haber ve yazılar için yapılan yorumların otomatik filtrelemesi için yapılacak olan bir projenin ön çalışmasıdır. Veri tabanımızda 1 milyonun üzerinde haber ve yorum bulunmaktadır. Elimizdeki verilerin yoğunluğundan dolayı deney seti olarak 44 farklı konuda yazılmış 15.064 adet gazete haberi ve makalesine yapılan 30.677 adet yorum kullanılmıştır. Literatürde yapılan sınıflandırma tabanlı yaklaşımlardan farklı olarak önerilen düzensizlik tabanlı yöntem de, yüksek hafıza gerekliliği ve yüksek hesaplama karmaşıklığına gerek kalmadan hızlı ve yüksek başarımda sonuçlar elde edilmiştir.

ABSTRACT

This is the preliminary work for a project which will be filtering comments made on news and papers automatically. Our database has over 1 million news and comments. Due to the intensity of our data, 30.677 comments made on 15.064 articles on 44 different categories are used as experimental data. Proposed anomaly based method have been obtained fast and high accuracy results without the high storage requirement and high computational complexity with respect to other classification based methods on literature.

1. GİRİŞ

Günümüzde metinlerin sınıflandırılması ya da istenmeyen e-postaların engellenmesi[1,2] ile ilgili olarak birçok çalışmalar yapılmaktadır ve yapılmaya devam edilecektir. Özellikle haber metinlerinin sınıflandırılması ve kategorilendirilmesi [3,4], yazar tanıma [5,6], web sınıflandırılması ve madenciliği [7] gibi birçok problemin temelinde metinlerin içeriklerinden elde edilmiş özellikler kullanılarak yapılan sınıflandırma temelli yaklaşımlar vardır. Bu tür yaklaşımlarda içerikten elde edilen bilginin önemi büyüktür. Elde edilen özelliklerin metinde kullanılan dilin özellikleri ile ilişkili olduğu ve o dilin getirdiği kurallara uygun olduğu varsayımı bulunmaktadır. Bundan dolayı bu tür içerik tabanlı çalışmalarda imla ve yazım kuralları öncelikli olarak ele alınır ve değerlendirilir.

Çalışmalar incelendiğinde makaleler üzerine yapılan yorumlar için yeterli akademik çalışma yapılmadığı görülmektedir. Bunun sebebi de yapılan yorumların genelde imla ve yazım kuralları ile örtüşmemesinden kaynaklanmaktadır.

Biz bu çalışmamızla akademik yazıma özellikle iki konuda katkı sağlamaktayız. İlki Türkçe metinler üzerinde yazım kurallarına ve noktalama işaretlerine uygun olmayan metinler için bir ayıklama ve ön işleme kuralları geliştirmek. İkincisi ise literatürde network alanında kullanılan aykırılık tabanlı yaklaşımı iyileştirerek son kullanıcı tarafından yazılmış yorumların otomatik olarak yayınlanıp yayınlanamayacağına karar veren bir karar destek sistemi hazırlamak.

Karşılaşılan en büyük problemlerden birisi, bilindiği üzere son kullanıcı yorumları, normal makale ya da haber yazılarından farklı olarak yazım ve imla kuralları ile genel olarak uyumlu değildir. Yazılan bu yorumlardan anlamlı bilginin elde edilmesi başlı başına bir problem olarak ele alınmaktadır. Literatürde geçen yayınlar incelendiğinde ise genelde yazım ve imla kuralları ile uyumlu olan haber ve makaleler üzerinde çalışıldığı görülmektedir. Bu çalışmamız bu konuda diğer çalışmalardan farklılık göstermektedir.

Karşılaşılan bir başka problem ise, yorumda bulunanların herhangi bir habere yaptıkları yoruma o anki ruh haletlerini ya da tutumlarını da katmaları ve bu tavrı sergilerken kullanmış oldukları kötü üslup ve argo tabirlerin bulunması ve yoruma etkisinin hesaplanmasının güç olmasıdır. Yorumun tamamını etkileyen kelimelerin ya da ilgili cümlenin çıkarılması, yorumun engellenmesi üzerinde çalışılması gereken bir başka problem olarak karşımıza çıkmaktadır. Bir gazetenin haber editörü olmanız durumunda haberlere yapılan her bir yorumun tek tek incelenip yorumun yayınlanıp yayınlanmayacağı ya da yorumun bazı yerlerinin değiştirilerek yayınlanmasına karar verme her zaman büyük bir iş yükü oluşturmaktadır. Bizim çalışmamız, bu tür aykırılıkları bularak yapılan yorumun engellenmesine ya da yayınlanmasına olanak sağlamaya yarayan bir karar destek sistemi olarak kullanılacaktır.

Aykırılık ya da farklılık tabanlı yaklaşımlar 19.yy.'dan bu yana istatistiksel olarak kullanılmaktadır [8]. Bu tür yaklaşımlar gürültüden arındırma [9] ve gürültü düzeltme [10] yöntemleri ile ilişkili gibi gözükse de aslında birbirinden farklı yaklaşımlardır.

Aykırılık ya da farklılık tabanlı yaklaşımların günümüzde sıkça kullanılmasının nedeni çeşitli alanlarda kolaylıkla uygulayabilme özelliğinden kaynaklanmaktadır. Özellikle sağlık alanında MRI görüntülerinden kötü huylu tümörlerin bulunması [11], ele geçirilmiş bir bilgisayardan gönderilen rastgele network paketlerinin engellenmesi [12], kredi kartı ya da kimlik hırsızlıklarının tespiti[13], terör faaliyetlerinin ortaya çıkarılması [14], bilgi çıkarımı [15, 16], sızmayı ve izinsiz girişlerin engellenmesi [17] ve sensor bozukluklarının tespiti [18] gibi konular örnek gösterilebilir. Veri madenciliği ya da bilgi çıkarımı alanlarında ise henüz yeni yeni kullanılmaya başlamaktadır.

Bu çalışma TÜBİTAK tarafından desteklenmektedir.
978-1-4673-0056-8/12/$26.00 ©2012 IEEE

2. VERİ KÜMESİ ve ÖZELLİKLERİ

Türkçe'miz sondan eklemeli bir dil olduğundan, kelimeler genel olarak *kelime kökü {+yapım ek(ler)i +{çekim ek(ler)i}}* yapısındadır. Yapım ekleri kelimenin anlamsal bilgisini ve/ya kelimenin türünü değiştirebilmektedir. Çekim ekleri ise sözcüğün anlamsal bilgisi üzerinde bir etkisi yoktur. Bu çalışmanın zorluğu diğer metin ve yazar sınıflandırmalarından farklı olarak, uyulması gereken birçok imla kuralı ve noktalama işareti kuralları yorum yapan kişiler tarafından dikkate alınmamasıdır. Bundan dolayı Zemberek gibi metin derleyicileri tam olarak doğru bilgiler ya da özellikler üretememektedir. *(Örnek: "vallah bende bilmiyommmmmmm
"; "Hayaller kurun Tabiiiiii ALEX ortya çkncy kadar.. ;)" vb…)*

2.1. Metin Derleyici

Literatürde metin ve yazar sınıflandırma konularında yapılan çalışmaların büyük çoğunluğu dili düzgün, noktalama işaretlerine ve imla kurallarına uygun yazılmış olan haber ve makale yazıları üzerinedir. Yaptığımız çalışma ise Türkçe haberler ve makaleler için yapılan yorumlar üzerindedir. Türkçe imla kuralları ve noktalama işaretleri öncelikli olarak dikkate alınmıştır. Fakat yorum metinleri çoğu zaman Türkçe imla kurallarına ve noktalama işaretlerine uygun değildirler. Çalışmamızda imla kurallarına uyan kelime oranının %30'u geçmediği görülmüştür. Bundan dolayı geliştirilen yöntemin önceliklerinden biri, imla ve yazım kurallarına uymayan kelime öbeklerini bularak yeni kurallar geliştirmek ve elimizdeki sözlükte bulunan kelimeye en yakın sözcüğü bularak yer değiştirmek şeklinde olacaktır.

Metin derleyicimizin diğer Türkçe metin derleyicileri gibi gerçeklediği bazı özellikler aşağıda sunulduğu gibidir:

- Girilen yorum metninin kelimeleri kök ve eklerine ayrılır.
- Eşleştirilemeyen kelimeler belirli kurallara göre sözlüğümüzdeki kelimeler ile eşleştirilmeye çalışılır, eşleştirilemeyen kelimeler şüpheli kelime sözlüğüne eklenir.
- Tüm kelimeler ve dokümanlar ile özellikleri MS SQL 2008'de oluşturulan veri tabanımızda kayıt altına alınır (yaklaşık *4 GB civarındadır*).

Yorumlarda, normal haber ya da makale metinlerinden farklı olarak çoğu imla kuralına uyulmamasından dolayı kelimeleri ayıklamak için bazı kurallar oluşturulmuştur. Metin derleyicimizin ele aldığı kurallardan bazıları aşağıdadır:

- *Kural 1:* Herhangi bir harf yerine herhangi bir noktalama işareti içeren kelimenin sözlükten uygun kelimeler ile eşleştirilmesi *(Örnek: nasıl bir yazı yazıyorsun eş.ek; n-yakınlık algoritması kullanılmıştır).*
- *Kural 2:* İçinde hiç sesli harf içermeyen kelimelerin sözlükten eşlenmesi, eğer uygun kelime bulunur ise yer değiştirmesi, bulunamaması durumunda şüpheli kelimeler tablosuna eklenmesi *(Örnek: ben bu gltsry gibi tkm hiç görmedim)*
- *Kural 3:* Türkçe'mizde yan yana iki sesli harf kullanılamaz kuralından yola çıkılarak; aynı sesli harfin iki ve ikiden fazla yan yana kullanılması durumunda sadece ilkinin kullanılması sessiz harflerde ise en fazla ikiye kadar izin verilmesi *(Örnek: Eveeetttt, hayırrrr, Neeeedeeeennn …).*
- *Kural 4:* Üç nokta haricinde kullanılan tüm iki ve ikiden fazla yan yana kullanılan noktalama işaretlerinin silinmesi.
- Kural 5: Sözlüğümüzle eşleşmeyen kelimeler üzerinde en yakın kelime için Türkçe-İngilizce harf ya da tam tersi yönde harf değişikliği yapılarak eşleştirme yapılması *(Örnek: ç↔c, ş↔s, ı↔i, ğ↔g, ö↔o, ü↔u).*
- Kural 6: Argo ve küfür kelimeleri sözlüğümüz ile eşleşen kelimelerin bulunması ve etiketlenmesi.

Yukarıdakiler gibi daha birçok kural *C#.Net* ve javada geliştirilmekte ve kütüphaneleri oluşturulmaya devam etmektedir. Şüpheli kelimeler tablomuzda yer alan kelimeleri çözümlemek için yeni kütüphaneler yazılmaktadır[1].

2.2. Yorum Özellikleri

Bu çalışma kapsamında Türkiye'deki popüler gazetelerden 15.064 adet makaleye yapılmış olan 30.677 adet yorum ele alınmıştır. Elde edilen makaleler ilk seviyeden 44 farklı konudan alınmıştır. Bazı konularda makale olmasına rağmen yorum yapılmadığından ya da yorum sayısı az olduğundan kategori sayısı 12 + 1 (diğer) ile sınırlandırılmıştır. Alt seviyelere (kategorilere) bu çalışmada inilmemiştir. Tablo 1'de kategoriler, makale sayıları ve bu makalelere yapılan yorum miktarları gösterilmektedir.

Tablo 1: Bir tablo örneği

Kategori	Haber Sayısı	Yorum Sayısı
Ana Sayfa	2.912	3.534
Son Dakika	2.672	4.176
Magazin	1.277	4.277
Ekonomi	569	873
Günaydın	317	970
Kültür Sanat	596	872
Politika	456	1.075
Yemekler	17	88
Spor	2.436	8.309
Teknoloji	1.927	3.572
Üçüncü Sayfa	51	199
Yazarlar	618	987
Diğer	1.216	1.745

Ayıklanan yukarıdaki yorum metinleri için,

- Yayınlanmış yorumlardan 10.614 adet kelime kökü
- Filtrelenmiş/engellenmiş yorumlardan 5.283 kelime kökü

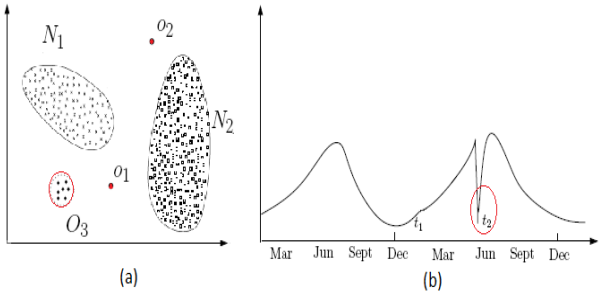
elde edilmiştir. Elde edilen kelime köklerinden 4.845 âdeti ortak olup 438 âdeti sadece filtrelenmiş yorumlara aittir. Bu elde edilen ayrık kelimeler daha önceden belirtildiği üzere şüpheli kelimeler listemize eklenerek gözden geçirilmektedir. Uygun olanlar argo terimler sözlüğümüze de eklenmeye devam etmektedir.

3. YÖNTEM

Düzensizlik tabanlı yaklaşımlar temelde 3 ana başlıkta toplanabilir. *Nokta düzensizlikler*, verinin dağılımına bakılarak belirli bir hata payına rağmen veri kümesi sınırları dışında kalan örneklerdir (Şekil 1.a). *Koşullu ve içeriksel düzensizlikler* herhangi bir örnek belirli bir özellik veya içerik için düzensizlik oluşturması durumudur. Çoğu olayda zaman faktörü dikkate alınarak ortaya çıkmaktadır. Kendi içinde

[1] *Akademik araştırmalar da kullanmak ve referans göstermek şartı ile dosyalar http://www.ugurayan.com sitesinden ya da mail ile istenebilir.*

davranışsal ya da içeriksel şekilde ikiye ayrılır (Şekil 1.b). Son tip olan *işbirlikçi düzensizlikler*de ise birden fazla örneğin genel veriye göre düzensizlik oluşturmasına dayalı durumlardır. Bu tür düzensizlikler bireysel olarak içinde bulunduğu veri kümesi içinde düzensizlik göstermese de bu düzensizliklerin işbirliğinden bir düzensizlik oluşmaktadır (Şekil 2). Buna en güzel örneklerden birisi web tabanlı saldırılar olabilir. Bu tür saldırılarda *buffer-overflow, ssh, ftp* gibi birden fazla özelliğin etkileşimi ve birleşimi ile saldırı oluşmaktadır.



Şekil 1: Nokta ve koşullu düzensizlik örnekleri

3.1. Algoritma

Önerilen yöntem, haber ve makaleler gibi imla ve yazım kurallarına uygun olan metinler yerine, çoğu kelime ve metin içeriğinin yazım kurallarına uygun olmayan yorumlar üzerinde yapılmaktadır. Bundan dolayı metinden bilgi çıkarımı (information extraction) ve ön işleme (preprocessing) adımları algoritmamızın önemli bir kısmını içermektedir. Diğer metin ayrıştırıcıların doğru çalıştığı birçok kural burada geçerli olmamaktadır. Bundan dolayı metinleri olabildiğince budayıp uygun formata dönüştürmek gerekmektedir.

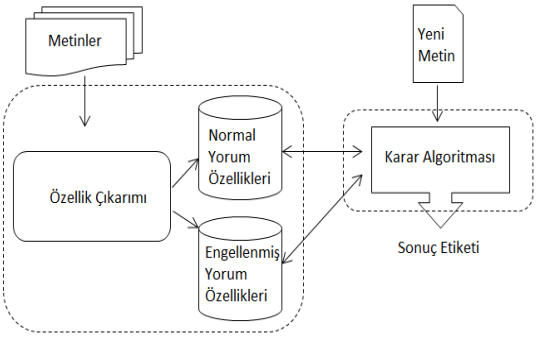
Algoritmamızın ön işleme adımında sırasıyla aşağıdaki işlemler uygulanmaktadır:

- Yorum metninden metin ayıklayıcı yardımı ile kelimeleri ayıklamaya çalış
- Ayıklanmış kelimelerin köklerini bul
- Kökleri bulunmayan kelimeler için geliştirilen budama yöntemlerinden sırasıyla
  - noktalama işareti içeren kelimeler için sırasıyla Kural 1-2 uygulanır
  - Kural 1-2 geçerli olmadığı durumlar için tekrar noktalam işaretine göre kelimeleri ayıkla ve kök bul
  - Kural 3-5 sırasıyla uygula
  - kelimenin kökü bulunamaması durumda şüpheli kelimeler sözlüğüne ekle
- budanan kelimelerin köklerini bul
- kökleri bulunamayan kelimeler var ise Kural 6 kapsamında oluşturulan argo terimler sözlüğümüz ile karşılaştır.

İstatistiksel açıdan ele alındığında verinin dağılımı,

$$D = (1-\alpha)M + \alpha A \tag{1}$$

şeklinde ifade edilirse, burada M verisetinden elde edilmiş olasılıksal çoğunluk (*majority*) dağılımını, A ise olasılıksal düzensizlik (*anomalous*) dağılımını göstermektedir. Genel düzensizlik tabanlı yaklaşımlarda başlangıçta tüm veri noktaları çoğunluk veri setine (M) ait olduğu varsayılmaktadır. Herhangi bir t zamanında $log\ D$ olasılığı $L_t(D)$ olmak şartı ile $\Delta = |L_{t+1}(D) - L_t(D)|$ şeklinde ifade edildiğinde, elde edilen bu fark değeri belirli bir eşik değeri (τ) aşar ise o zaman düzensizlik ortaya çıkmış olur şeklinde ifade edilir.



Şekil 2: Yorum yayınlama karar mekanizması.

Her hangi bir zamandaki log olasılığını bir model (*naive Bayes, maximum entropy, vb...*) ile ifade etmek için denklem 2 baz alınmaktadır.

$$L_t(D) = \prod_{i=1}^{N} P_D(x_i) = \left(\alpha^{|M_t|} \prod_{x_i \epsilon M_t} P_{M_t}(x_i)\right). \left((1-\alpha)^{|A_t|} \prod_{x_i \epsilon A_t} P_{A_t}(x_i)\right) \tag{2}$$

bu genel tanımlanan denklemi aşağıdaki denkleme indirgeyerek logaritmik modeli oluşturduk.

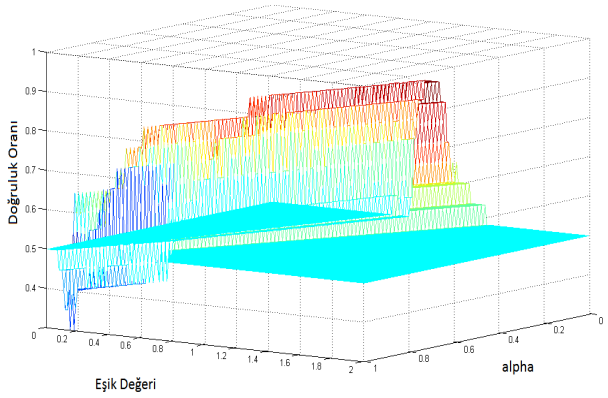
$$LL_t(D) = \alpha . \frac{1}{s(x_i \epsilon M_t)} \left(\sum_{x_i \epsilon M_t} \log P_{M_t}(x_i)\right) + (1-\alpha) \frac{1}{s(x_i \epsilon A_t)} \left(\sum_{x_i \epsilon A_t} \log P_{A_t}(x_i)\right) \tag{3}$$

burda α çoğunluk dağılımının önem olasılık katsayısı, $P_{M_t}(x_i)$ ise t dokümanının x_i'nci çoğunluk özelliğinin genel ortalama dağılımından ne kadar saptığı bilgisini, $s(x_i \epsilon M_t)$ ile $s(x_i \epsilon A_t)$ ise dökümanın sırasıyla çoğunluk ve düzensizlik özelliği sayısını, $P_{A_t}(x_i)$ ise düzensizlik oran bilgisini göstermektedir. $L_t(D) > \tau$ veya $LL_t(D) > \tau$ olan durumlarda yorum filtrelenecektir.

4. SONUÇLAR

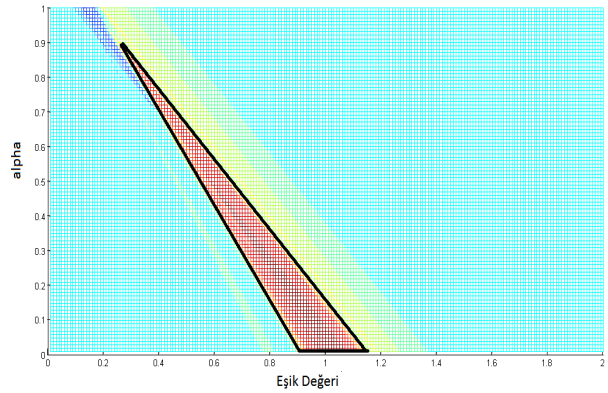
Önerilen yöntemin en iyilem aralığını bulmak için açgözlü (greedy–search) arama ile alpha ve eşik değerleri hesaplanmıştır (Şekil 3).

Yorum filtreleme için önerilen metodun en iyi parametrik bilgilerini elde etmek için Şekil 4 incelendiğinde etrafı çizili alanın doğruluk oranı %90 ve üzerinde olduğu görülmektedir.



Şekil 3: Eşik değeri ve çoğunluk dağılımının önem olasılık katsayısı değerlerine bağlı 3 boyutlu doğruluk oranı.

Bu çalışmada iki temel kavram üzerinde çalışılmıştır. Bunlardan ilki, tam olarak imla ve yazım kurallarına uyumlu olmayan metinlerden kelime ayıklama işlemi ve anlamlı bilgi çıkarımı için kurallar oluşturmak, ikincisi ise hızlı ve doğru bir algoritma geliştirmektir. Bunun için literatürde metin sınıflandırmada kullanılmayan düzensizlik tabanlı yaklaşım kullanılmıştır.



Şekil 4: Eşik değeri ve çoğunluk dağılımının önem olasılık katsayısı değerlerine bağlı 2 boyutlu doğruluk oranı.

Elde edilen sonuçlar yorum sınıflandırmada uygun eşik değeri ve çoğunluk dağılımının önem olasılık katsayısı (α) verildiğinde yorumları filtrelemede doğruluk oranının %90'ı geçtiği en yüksek oran olarak da %98.7 olarak hesaplanmıştır.

5. KAYNAKÇA

Luo, Q., Liu, B., Yan, J., He, Z., "Design and Implement a Rule-Based Spam Filtering System Using Neural Network", *Proc. Int. Computational and Information Sciences (ICCIS) Conf,* 398-401, 2011.

Hayat, M. Z.,Basiri, J., Seyedhossein, L., Shakery, A., "Content-based concept drift detection for Email spam filtering", *Proc. 5th Int. Telecommunications (IST) Symp,* 531-536, 2010.

Joachims,T., "Retrospective on Transductive Inference for Text Classification using Support Vector Machines. *Proceedings of the International Conference on Machine Learning (ICML),* 1999.

Li,Y.,Chung, S.M., Holt, J. D., "Text document clustering based on frequent word meaning sequences", *Data Knowl. Eng.,* 64(1):381–404, 2008.

Diri, B., Kaban, Z.,"Genre and author detection in Turkish texts using Artificial Immune Recognition Systems", *IEEE 16th Signal Processing Communication and Applications Conference Proc.*, 2008.

Amasyalı, M.F., Diri, B., "Automatic Turkish Text Categorization in Terms of Author, Genre and Gender", *11th International Conference on Applications of Natural Language to Information Systems-NLDB 2006*, Austria, LNCS Volume 3999, 2006.

Langville, A. N., Meyer, C. D.,"A Survey of Eigenvector Methods for Web Information Retrieval", SIAM Review, Society for Industrial and Applied Mathematics, 47, pp. 135-161, 2005.

Edgeworth, F. Y., "On discordant observations", *Philosophical Magazine* 23, 5, 364-375, 1887.

Teng, H.,Chen, K., and Lu, S., "Adaptivereal-time anomaly detection using inductively generated sequential patterns", *In Proceedings of IEEE Computer Society Symposium on Research in Security and Privacy. IEEE Computer Society Press,* 278-284, 1990.

Rousseeuw, P. J. And Leroy, A. M., *Robust regression and outlier detection,* John Wiley & Sons, Inc., New York, NY, USA, 1987.

Spence, C.,Parra, L., and Sajda, P., "Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model", *In Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis.* IEEE Computer Society, Washington, DC, USA, 3, 2001.

Kumar, V., "Parallel and distributed computing for cyber security", *Distributed Systems Online,* IEEE 6, 10, 2005.

Aleskerov, E., Freisleben, B., and Rao, B., "Cardwatch: A neural network based database mining system for credit card fraud detection", *In Proceedings of IEEE Computational Intelligence for Financial Engineering,* 220-226, 1997.

Zackrison, J.L., "La Violencia in Columbia: Annomaly in Terrorism", The Jornal of Conflict Studies, Vol.9(4), 1989.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark,R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E.,"PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals", *Circulation,* vol.101, 23, e215-e220, 2000.

Eskin, E., "Anomaly detection over noisy data using learned probability distributions", *In Proceedings of the Seventeenth International Conference on Machine Learning,* Morgan Kaufmann PublishersInc., 255-262, 2000.

Barbara, D., Couto, J., Jajodia, S., and Wu, N., "Detecting novel network intrusions using bayes estimators", *In Proceedings of the First SIAM International Conference on Data Mining*, 2001.

Fujimaki, R.,Yairi, T., and Machida, K., "An approach to space craft anomaly detection problem using kernel feature space", *In Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM Press, New York, NY, USA, 401-410, 2005.