

Gene 3E:

Genetik Veri Madenciliği için Yeni Analiz Aracı

^{1,3} Erdal Coşgun , ^{2,3} Aydın Kaya

Ali Seydi Keçeli^{2,3}, Yiğitcan Aksarı^{2,3}, Prof.Dr.Ergun Karaağaoğlu¹

¹ Biyoistatistik AD, Tıp Fakültesi, Hacettepe Üniversitesi, Ankara

² Bilgisayar Mühendisliği, Mühendislik Fakültesi, Hacettepe Üniversitesi, Ankara

³ Biyoinformatik ve Veri Madenciliği Çalışma Grubu, Ankara
<http://www.hacettepebiyoinformatik.org/>

Abstract: Genetic researches which began structured and comprehensive way with HapMap project, continuing in full swing in recent years. The size of the data from these studies remain significant. Provision for investments to be made possible through the analysis of these data effectively. Incorrect or biased results can be lead to serious problems. This study describes "Gene 3E" program in detail which was developed for use an important part of the available data in genetic: DNA-microarray gene expression and Single Nucleotide Polymorphisms (SNP) . "Gene 3E" is a tool that contains all the necessary stages for multi-class- classification problems and determination of candidate "gene or SNPs" problems for genetic data. Gene 3E combines Java user interface and [R] software which allows the "hybrid-method" usage that has many advantages for user benefits. Tool basically contains "generalization, dimensionality reduction, clustering and classification" sub-analysis groups.

Key Words: Bioinformatics, Data Mining, Microarray Gene Expression Data, Classification, Clustering

Özet: Hap-Map projesi ile planlı ve kapsamlı şekilde başlayan genetik araştırmalar son yıllarda tüm hızıyla devam etmektedir. Bu çalışmalardan elde edilen verilerin büyüklüğü ise ciddi boyutlardadır. Yapılan yatırımların karşılığının alınması ise bu verilerin etkin şekilde analizi ile mümkündür. Yanlış ya da yanlış sonuçlar ciddi sorunlara yol açabilmektedir. Bu çalışma mevcut genetik verilerin önemli bir kısmını oluşturan DNA-mikroarray gen ekspresyon ve Tek Nükleotid Polimorfizmi (TNP) verilerinin analizinde kullanılmak üzere geliştirilen "Gene 3E" programını detaylı olarak anlatmaktadır. "Gene 3E" çok boyutlu genetik verilerin sınıflama problemleri ile aday "gen ya da TNP"lerin belirlenmesi için gerekli tüm aşamaları içeren bir araçtır. Java kullanıcı ara yüzünü ve [R] yazılımını birleştiren araç, kullanıcılara ciddi avantajlar getiren "hibrit-yöntem" kullanımına olanak tanımaktadır. Araç temel olarak "genelleştirme, boyut indirgeme, kümeleme ve sınıflama" alt analiz gruplarını içermektedir.

Anahtar Kelimeler: Biyoinformatik, Veri Madenciliği, Mikroarray Gen Ekspresyon Verisi, Sınıflama, Kümeleme

1. Giriş

Veri madenciliği (DM=Data Mining) ve Makina öğrenmesi (ML=Machine Learning) yaklaşımlarının genetik araştırmalarda kullanımı son yıllarda giderek artmaktadır [1]. Araştırmalar sonucunda yüzlerce ya da milyonlarca genetik etken (mikroarray gen ekspresyon verileri) ya da Tek Nükleotid Polimorfizmi (TNP¹: Single-Nucleotide Polymorphism) içeren genetik veri matrisleri elde edilir. Ancak sadece verilerin elde edilmesi yeterli olmamaktadır. Bu verileri doğru ve yansız biçimde analiz edebilecek istatistiksel modellere ihtiyaç duyulmaktadır.

DM ve ML yöntemlerinin yaygın şekilde kullanıldığı klinik çalışmalar DNA mikroarray² çalışmalarıdır[2,3,4,5,6,7]. İnsan genomunda yaklaşık 40.000 gen bulunduğu göz önüne alındığında bu kadar çok genin tek tek analizi mümkün değildir. Fakat günümüzde, geliştirilen otomasyona dayalı sistemlerle çok fazla sayıdaki genin aynı anda analizi mümkün hale gelmiştir. Bu analizlerin önemli bir bölümünü genlerin sınıflanması ve önemli genlerin bulunması oluşturmaktadır. Ancak gen araştırmaları hem maliyeti hem de ölçümlerin tekrar edilmesindeki zorluklar nedeniyle çok fazla hasta üzerinde yapılamamaktadır. Bu nedenle az sayıdaki hastaya ait yüzlerce gen verisi üzerinden bazı yorumlara ulaşılacak durumda kalınmıştır. Klasik istatistiksel yöntemler (lojistik regresyon, varyans analizi, doğrusal regresyon analizi) bu tip verileri açıklamakta sorun yaşamaktadır. Son dönemde bu tip veriler üzerinde Destek Vektör Makinaları (SVM: Support Vector Machines), Yapay Sinir Ağları (NN: Neural Networks), Karar Ağaçları, Naive Bayes, Random Forest gibi birçok veri madenciliği yöntemi denenmiş ve iyi sonuçlar elde edilmiştir[2,3,4,5,6,7].

Son yıllarda yapılan çalışmalarda ise mikroarray veriler dışında TNP'ler gibi genotip verilerin DM ve ML yöntemleri ile analizi eklenmiştir[8,9,10,11]. Bu verilerin mikroarray verilerden farkı ise veri boyutlarının çok daha fazla olmasıdır. Şu ana kadarki en önemli araştırma GWAS (Genome-Wide Association Study) çalışmasıdır. Bu çalışma belirli türlere ait bireylerin tüm ya da birçok geninin bireyden bireye ne kadar değiştiğini anlamayı hedefleyen bir çalışmadır. Bu çalışma sayesinde bir kişiye ait milyonlarca TNP bilgisini elde etmek mümkün olmuştur. TNP'lerden elde edilecek her bilgi hastalık tedavisinde önemli rol oynamaktadır. TNP analizleri genel olarak "aday TNP" bulmaya odaklanan yöntemler üzerine yoğunlaşmıştır. Bu yöntemler sayesinde milyonlarca TNP arasında belirli bir hastalığa etki eden önemli TNP'ler belirlenebilmektedir. Belirlenen TNP'lere uygulanacak tedaviler insanlık için kritik önem taşımaktadır. TNP verileri ile birlikte DNA dizilerinin (sekans) analizi de genetik araştırmalar için önemlidir. Ancak bu veri türleri için veri madenciliği-makine öğrenmesi yöntemlerinin kullanımı henüz çok temel düzeydedir. Özellikle "Gelecek Nesil(Next Generation) Dizileme" verileri çok yeni araştırma alanlarındadır.

Biyoinformatik çalışmalarda çok boyutlu genetik verileri (mikroarray gen ekspresyon verisi, Tek Nükleotid Polimorfizmi (TNP) analiz etmek en önemli problemlerdendir. Bu tip verileri analiz etmek için son yıllarda birçok analiz aracı araç tasarlanmıştır(Bioconductor, SamTools, Orange vb.). Ancak bu analiz araçları genellikle kısıtlı sayı ve türde yöntemleri içermektedir. Bununla birlikte yeni yöntemlerin bu araçlara dahil edilmesi de zaman almaktadır. Çünkü araçlar genellikle kendilerine özel kodlama platformları kullanmayı tercih etmektedirler.

Aynı zamanda tıp ya da biyoloji alt yapısı olan araştırmacılar araçlar hakkında kapsamlı bir eğitimi almak zorunda kalmaktadırlar. Araç hakkında yeterli bilgisi olmayan kullanıcıların yaptığı analizler de yanlış bulgulara ulaşılmasına sebep olmaktadır. Sık yapılan bazı hatalar şunlardır: kullanılacak yöntemlerin parametrelerinin optimizasyonun yapılmaması, yöntemlerin yanlış sıralama ile kullanılması, genetik verilerin boyut indirgemeden analiz edilmesi. Kullanılması gereken tüm yöntem gruplarını içeren, kullanımı görece olarak daha kolay, güncelleme olanağı fazla olan bir "biyoinformatik" analiz aracına ihtiyaç duyulmaktadır. Bu çalışmada sunulan "Gene 3E" analiz aracı, açık kaynak kodlu ve ücretsiz bir yazılım olan [R]'a ait paketleri kullanmaktadır. Şimdiye kadar [R] yazılımı içerisinde birçok farklı analiz aracı geliştirilmiştir. Ancak "Gene 3E" bunlar içinde biyoinformatik sorunların çözümünde hibrit yöntem kullanımına olanak tanıyan nadir örneklerdendir.

[R] yazılımı içinde en sık kullanılan veri madenciliği aracı "rattle"dır. Bu araç [R] içerisindeki veri madenciliği paketlerini (kayıp gözlem analizi, dönüşüm, karar ağacı yöntemleri) bir araya getirir. Ancak yeni kullanıcılar için kullanımı zordur.[12]. Son yıllarda [R] yazılımı kullanılarak oluşturulan bir başka veri madenciliği aracı ise "miRD"(microRNA detection). Bu araç özellikle önemli, pre-microRNA'ların belirlenmesinde makine öğrenmesi yöntemlerinden SVM kullanmaktadır.[13] Marc Johannes ve ark. Çalışmalarında "pahClass" adlı [R] aracını önermişlerdir. Bu paket ise farklı birçok paketi farklı sıralamalar ile kullanamaya olanak tanımaktadır. Genel olarak SVM tabanlı bir yaklaşım içermektedir. [14] Ancak bu araçlardan hiçbirisi genetik verilere özel geliştirilmemiştir. Bu nedenle bu tip verilere özel yaklaşımları içermemektedir.

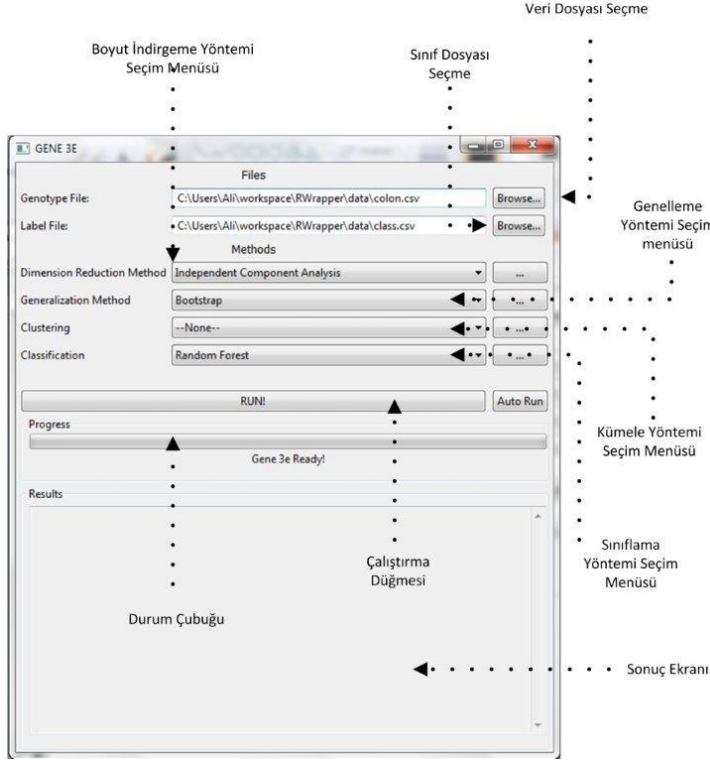
2. Amaç

Bu çalışmada amaçlanan "yüksek boyutlu genetik" verilerde (gen ekspresyon, TNP) sınıflama ve önemli genetik faktörlerin bulunması problemleri için gerekli ve destekleyici yöntemleri içeren bir java analiz aracı geliştirmektir. (Şekil-1) Geliştirdiğimiz araç dört alt analiz grubuna sahiptir. Bunlar:

1. Boyut İndirgeme
2. Genelleştirme Yöntemi
3. Sınıflama

4. Kümeleme

“Gene 3E” farklı yöntemin hibrit şekilde kullanılabilmesine olanak tanıyacak şekilde tasarlanmıştır. Bu kombinasyonların çok sınıflı sınıflama problemlerindeki başarımının karşılaştırılması için **Doğruluk (Accuracy)** kriteri kullanılmıştır. Ancak eğer kullanıcı iki sınıflı-sınıflama problemi için aracı kullanıyorsa beş farklı kriterin sonucu rapor edilmektedir. Bunlar: **Matthews Korelasyon katsayısı, Recall , Precision, F-Ölçüsü**



Şekil 1: “Gene 3E” kullanıcı ara yüzü

3. Gereç ve Yöntemler

Genetik veriler deneyler sonunda elde edilip depolandıktan sonra en önemli aşama biyolojik örüntülerin gözlenen fenotip üzerine etkisini belirlemektir. Genetik verilerin analizi bilinen klasik istatistiksel yöntemlerin (t-testleri, Varyans analizleri) ötesinde veri madenciliği yöntemlerini kullanmayı gerektirmektedir. Bunun en önemli nedeni, VM yöntemlerinin çok yüksek boyutlu verilerde ‘doğrusal olmayan’ ilişkileri belirlemede avantaja sahip olmasıdır. Son yıllarda işte bu avantajlı yöntemlerin kullanılabilmesi için birçoğu açık kaynak kodlu program geliştirilmiştir. Bunlardan en önemli [R] yazılımıdır.

3.1. Yöntemler

Tüm analizler “bootstrap” yaklaşımını genelleştirme yöntemini kullanmaktadır. Yerine koyarak seçilen çok sayıdaki örneklemin genetik verilerde çok daha tutarlı sonuçlar verdiği çalışma grubu tarafından 1 nolu kaynakta ispatlanmıştır. “Gene3E”, iki boyut indirgeme, üç sınıflama ve kümeleme yönteminin hibrit şekilde kullanımını desteklemektedir. Tablo 1.’de yöntemler ve kullandıkların [R] paketleri gösterilmiştir.

3.2. R Yazılımı ve Rserve

3.2.1. [R] Yazılımı

[R] , VM çalışan her araştırmacı için standart olan, açık kaynak kodlu ve ücretsiz bir yazılımdır. R programının mantığı her analiz için kullanılabilecek ‘analiz paketlerine’ sahip olmasıdır. Kullanıcılar yapmak istedikleri analize ait paketi indirdikten sonra ilgili paketin kullanım kılavuzuna göre analizlere devam ederler. R’in en önemli avantajı diğer paket programlara göre çok hızlı olmasıdır. Standart bir bilgisayar yapılandırması ile (örn:

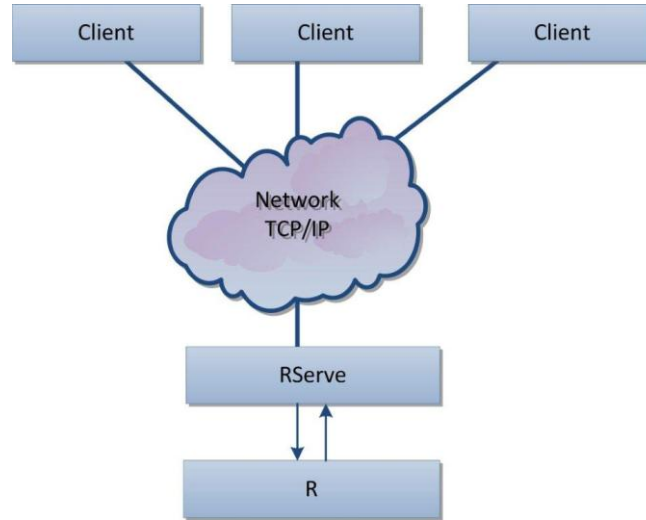
2GB RAM ve 2.13 GHz işlemci) çok büyük veri setleri kısa sürede analiz edilebilir. Bu programa ait tüm bilgiye bu bağlantıdan <http://www.r-project.org/>, yöntemlerin paketlerine ait tam listeye ise bu bağlantıdan ulaşılabilir : <http://cran.r-project.org/src/contrib/PACKAGES.html>. Bu çalışma kapsamındaki VM yöntemlerine ait R paketleri Tablo 1’de verilmiştir. Bu paketlerin dışında ilgili yöntemler için kullanılabilecek başka kaynak paketler de bulunabilir.

Tablo 1. Gene 3E’nin içerdiği yöntemler ve [R] paketleri

ANALİZ TÜRLERİ	YÖNTEM	[R] PAKETİ
BOYUT İNDİRGEME	<i>Bağımsız Bileşenler Analizi</i>	fastICA
	<i>Temel Bileşenler Analizi</i>	princomp- <i>{stat}</i>
SINIFLAMA	<i>SVM</i>	e1071
	<i>Naive Bayes</i>	e1071
	<i>Random Forest</i>	randomForest
KÜMELEME	<i>K-Means</i>	k-means- <i>{stat}</i>
	<i>K-Medoid</i>	cluster
	<i>Fuzzy C-Means</i>	FactoMineR
	<i>Kohonen Map</i>	som

3.2.2. RServe

RServe, Java ve C++ gibi diller ile geliştirilmiş yazılımlarda, R ile TCP/IP üzerinden bağlantı kurmak için kullanılan bir sunucu yazılımıdır [16]. RServe, uzaktan R oturumları açmaya olanak sağlar. Açılan her bağlantı, farklı bir çalışma alanı ile ilişkilendirilir ve R kütüphanelerinin istemci tarafından kullanılabilir (Şekil 2). Bu bağlantı yönteminin basitliği birçok işlevsel aracın gerçekleştirilmesine olanak sağlar. Bu sayede bir kullanıcı, R bilgisi olmadan veya R kodları üzerinde değişiklik yapmadan karmaşık betikleri işletebilir.



Şekil 2: İstemcilerin, TCP/IP üzerinden R/RServe ile iletişim mimarisi.

3.2.3 Uygulama

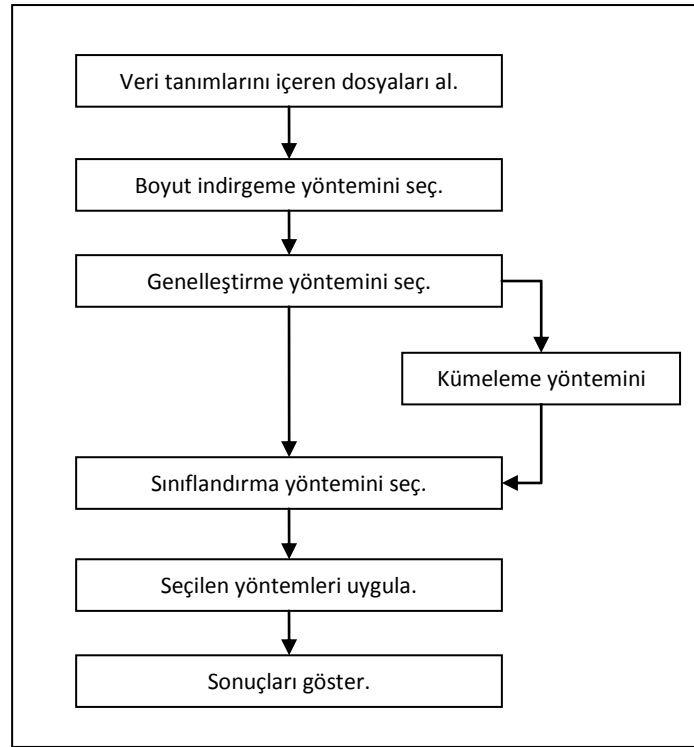
Gene3E, genetik araştırmalarda veri madenciliği yöntemlerini kullanmak üzere Java programlama ortamı kullanılarak geliştirilmiştir. Görsel arayüzün oluşturulmasında SWT (Standard Widget Toolkit) kullanılmıştır [17]. Arayüz sayesinde kullanıcı istatistiksel arka plandan soyutlanmakta ve çözümlemelerin R dilinde nasıl gerçekleştirildiğiyle ilgilenmeden kolayca yapabilmektedir.

Uygulama akışı, veriler girildikten sonra dört farklı adım seçiminden oluşmaktadır (Şekil 3). Bu adımlar boyut indirgeme, genelleştirme, kümeleme (isteğe bağlı) ve sınıflandırmadır. Her bir adım için farklı yöntemler seçilebilir. Bu yolla otuz farklı kombinasyonla analiz yapılabilmektedir.

Yukarıda listelenen yöntemlerin parametreleri, girdi olarak verilen genetik veriye göre özelleştirilebilir. Yöntemlerin en iyileştirilmiş parametrelerini bulmak önemli bir işidir. Analizler sonucunda, sonuç ekranındaki özet veriler değerlendirilerek kullanıcı kendi verisine uygun parametreleri belirleyebilir. Gene3E üzerinde tanımlanmış varsayılan parametreler Tablo 2’de yöntemler ve kullanıcı tarafından R paketleri üzerinde değiştirilebilecek parametrelerin listesi verilmiştir.

Tablo 2: Yöntemler ve kullanıcı tarafından değiştirilebilecek parametreleri.

	Yöntem	Parametreler					
Boyut İndirgeme	Independent Component Analysis (ICA)	Number of components	Algorithm type	Number of max. iter.	Tolerance		
	Principal Component Analysis (PCA)	Number of components					
Genelleştirme	Bootstrap	Number of variable	Bootstrap number				
Kümeleme	K-means	Centers	Max. iteration	Algorithm	Nstart		
	K-medoid(PAM)	K (centers)	Metric				
	Kohonen SOM	xdim	ydim				
	Fuzzy C-means	Centers	Max. iteration	Distance	Method	m	rate.par
Sınıflandırma	Support Vector Machine (SVM)	Cost	Tolerance	Kernel			
	Random Forest	Number of trees	Number of var.	Importance	Proximity		
	Naive Bayes	Laplace	NA action				



Şekil 3: Uygulamanın akış çizeneği.

4. Sonuç

Bu çalışmada önerilen analiz aracı genetik araştırmalara özeldir. Genetik verilerde sınıflama ve önemli genetik etkenlerin belirlenmesi analizinde “boyut indirgeme” zorunlu bir adımdır. Çünkü genler ya da TNP’ler gizli bazı ilişkilere sahiptir. Bu ilişkiler göz ardı edilerek yapılan analizler biyolojik açıdan yanlış sonuçlar verebilir. Bu nedenle biz de aracımızda bu adımı kullanıcılara ilk aşamada zorunlu olarak kullanılmaktayız. Yapılan birçok

çalışmada boyut indirgemenin analizlerde daha iyi sonuçlar verdiği gösterilmiştir[18,19,20]. Bu adımdan sonra kullanıcılar isterlerse “sadece sınıflama” ya da kümeleme analizinden sonra elde edilen küme merkezlerini kullanarak sınıflama analizleri yapabilmektedirler. Bu yaklaşım özellikle boyut indirgemenin yeterli olmadığı durumlarda kullanılmaktadır[1]. “Gene3E” sınıflama problemleri için çok önemli bir özelliğe sahiptir. Birçok analiz aracı hangi yöntem ya da yöntem grubunun daha iyi performans gösterdiği konusunda kullanıcılarına yol göstermemektedir. Ancak “Gene3E” iki sınıflı-sınıflama problemlerinde beş farklı performans ölçüsünün sonucunu aynı anda vererek kullanıcıların daha bilinçli seçimler yapmasını sağlamayı hedeflemektedir.

Çalışmadaki bir diğer önemli nokta ise genetik araştırmalarda çok sık kullanılan “çapraz geçerlilik (ÇG)” yöntemi yerine genelleştirme yöntemi olarak “bootstrap” yönteminin seçilmesidir. Bu seçimdeki en önemli sebep farklı genetik faktörlerin bulunmasında ÇG’nin zayıf kalmasıdır. Özellikle az sayıda birey üzerinde yapılan genetik deneylerde ÇG’de her aşamada analiz modellerini test etmek için dışarıda bırakılan kişi sayısı çok az olmaktadır. Örneğin, 20 kişilik bir deney için 10-parça ÇG’de her adımda 2 kişi ile model test edilmektedir. Bu sorunu çözmenin en etkili yolu ise “bootstrap” ile modelleri en az 100 en fazla 1000 defa tekrarlamaktır. [R] yazılımı sayesinde bu yaklaşım çok tutarlı şekilde yapılabilmektedir. En iyi performansı gösteren alt grupları raporlaması sayesinde özellikle “aday gen-TNP” bulunması analizlerinde önemli rol oynayacağını düşünmekteyiz.

Gene 3E’nin en önemli özelliklerinden birisi de hızlı olmasıdır. Intel Core2 Duo-2.20 GHz işlemci ve 4 GB yüklü belleğe sahip bir bilgisayarda, 100 kişiye ait 2000 gen’in ekspresyon verisi için 2 sınıflı-sınıflama analizi en fazla 2 dakika içerisinde analiz edilmektedir. Aynı şekilde 60 bireye ait 10000 TNP verisi için 6 dakikalık bir süreye ihtiyaç duyulmaktadır. Literatürde bu tip analizleri hibrit şekilde yapan herhangi bir örnek olmadığı için şu an için karşılaştırma yapmakta mümkün olamamaktadır. Sonraki çalışmalarımızda “Gene 3E” kapsamındaki yöntemlerin sayısının artırılması ve parametre optimizasyonu için kullanıcılara bazı seçenekler tanınması planlanmaktadır. Ayrıca ülkemizde genetik araştırma yapan bilim insanlarına aracımızı anlatmak ve tanıtmak kısa vadeli hedeflerimizdendir. Bununla birlikte sonuçların grafik sonuç gösterimi, analiz süresinin kısaltılması ve diğer özelliklerin geliştirilmesi Texas A&M-Commerce Üniversitesi, Bilgisayar Bilimleri bölümü desteği ile HPC (High Performance Computing) tesislerinde devam etmektedir.

Not:

Bu çalışma söz konusu analiz aracının ortaya çıkış çalışması olan [15] nolu kaynağın devamı olarak yazılmıştır. İlk çalışmada sadece aracın temel fonksiyonları ortaya konmuştur. Bu çalışma ile ilk çalışmaya yöntem ve performans kriteri bakımından ciddi eklemeler bulunmaktadır.

Sorumlu Yazar:

Ar.Gör.Erdal Coşgun

GSM: 536 424 11 30

E-posta : erdal.cosgun@hacettepe.edu.tr

<http://yunus.hacettepe.edu.tr/~erdal.cosgun>

KAYNAKÇA

- [1] Erdal COŞGUN, Nita Limdi, Christine W.Duarte, High Dimensional Pharmacogenetic Prediction of a Continuous Trait using Machine Learning Techniques with Application to Warfarin Dose Prediction in AfricanAmerican,Bioinformatics, 2011, 27,10,1384-1389
- [2] Erdal COŞGUN, Ergun Karağaoğlu, The New Hybrid Method for Classification of Patients by Gene Expression Profiling, In: Suh, Sang C.; Gurupur, Varadraj P.; Tanik, Murat M. (eds), Biomedical Engineering: Healthcare Systems, Technology and Techniques, Springer , 1st Edition., 2011, 255-265
- [3] Rudolph S. Parrish, Horace J.Spencer, Ping Xu, Distribution Modelling and Simulation of Gene Expression Data, Computational Statistics and Data Analysis, 2009
- [4] Vladimir Vapnik. Estimation of Dependences Based on Empirical Data [in Russian]. Nauka, Moscow, 1979. (English translation: Springer, New York, 1982)
- [5] Erdal Coşgun, Dr.Erdem KARABULUT, Prof.Dr. Ergun KARAĞAOĞLU, “Random Forest (RF) ve SVM ile Mikroarray Verilerde Gen Seçimi”, VI. Ulusal İstatistik Kongresi, 29 Nisan – 3 Mayıs 2009, Antalya, Türkiye

- [6] Arun Jagota, Microarray Data Analysis and Visualization, Bioinformatics By the Bay Press, Santa Cruz, 2001
- [7] Pablo Tamayo et al., Interpreting Patterns of Gene Expression with Self-Organizing Maps : Methods and Application to hematopoietic differentiation. Proc. Natl. Acad. Sci., 1999; Volume 96, 2907-2912
- [8] Wang M, Chen X, Zhang M, Zhu W, Cho K, Zhang H. 2009. Detecting significant single-nucleotide polymorphisms in a rheumatoid arthritis study using random forests. BMC Proc 3:S69.
- [9] Wu TT, Chen YF, Hastie T, Sobel E, Lange K. 2009. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics, 25:714–721.
- [10] Yang W, Gu CC. 2009. Selection of important variables by statistical learning in genome-wide association analysis. BMC Proc, 3:S70.
- [11] Tomita Y, Tomida S, Hasegawa Y, Suzuki Y, Shirakawa T, Kobayashi T, Honda H. 2004. Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. BMC Bioinformatics, 5:120
- [12] Graham J Williams (2009), Rattle: A Data Mining GUI for R, The R Journal Vol.12
- [13] Yuanwei Zhang et.al (2011), Prediction of Novel Pre-microRNAs with High Accuracy through Boosting and SVM, Bioinformatics (Published Online)
- [14] Marc Johannes et.al, (2011), pathClass: An R-Package for Integration of Pathway Knowledge into Support Vector Machines for Biomarker Discovery, Bioinformatics (Published Online)
- [15] Erdal Cosgun, Yiğitcan Aksarı, GENE 3E : A New Bioinformatics Tool For Genetic Data Mining, Society for Design and Process Conference, June 12– 16, 2011, Jeju, South Korea
- [16] <http://rosuda.org/Rserve/>
- [17] Steve Northover and Mike Wilson. 2004. Swt: The Standard Widget Toolkit, Volume 1 (First ed.). Addison-Wesley Professional
- [18] Su-In Lee, Serafim Batzoglou, An Application of Independent Component Analysis to Microarrays, Genome Biology, 2003; 4:R76Ka Yee Yeung, Mario Medvedovic and Roger E. Bumgarner, Clustering Gene Expression Data With Repeated Measurements, Genome Biology, 2003; 4:R74
- [19] International Journal of Innovative Computing, Information and Control ICIC International, Independent Component Analysis For Classification Of Remotely Sensed Images, 2006; Volume 2, Number, 31349-4198,
- [20] Breiman, Leo, "Random Forests". Machine Learning 45 (1): 2001; 5–32.