Gen İfade Verilerinde Eksik Değerleri Düzelten Kestirim Yöntemlerinin Karşılaştırılması

Comparison of Estimation Methods for Missing Value Imputation of Gene Expression Data

Ali Sarıkaş[1], Niyazi Odabaşıoğlu[2], Gökmen Altay[3]
[1]Elektronik ve Otomasyon Bölümü, Marmara Üniversitesi TBMYO, İstanbul, Türkiye
ali.sarikas@marmara.edu.tr
[2]Elektrik Elektronik Mühendisliği Bölümü, İstanbul Üniversitesi, İstanbul, Türkiye
niyazio@istanbul.edu.tr
[3]Biyomedikal Mühendisliği Bölümü, Bahçeşehir Üniversitesi, İstanbul, Türkiye
gokmen.altay@eng.bau.edu.tr

***Özetçe*—Mikrodizi gen ifade veri kümelerinin ön işleme sürecinin ilk basamağı, eksik değerlerin kontrolü ve düzeltilmesi işlemidir. Bu makalede, eksik değerlerin kontrolü ve düzeltilmesi için şimdiye kadar kullanılan en güvenilir ve güncel kestirim yöntemleri karşılaştırılmıştır. Eksik verilerin kontrolü ve düzeltilmesi işleminin sonraki ön işleme (normalizasyon, kalite kontrolü), farksal gen ifade veri analizi, sınıflandırma, kümeleme, yolak analizi, vs. aşamalarını da etkilediğinden yüksek doğrulukla yapılması çok önemlidir. Bu çalışmada en popüler 5 yöntemin (En yakın k-komşu, Bayesçi temel bileşen çözümlemesi, yerel en küçük kareler, ortalama ve ortanca) başarımlarını değerlendirmek için NRMSE değerleri kullanılmıştır. Yöntemlerin NRMSE değerleri incelendiğinde, yerel en küçük kareler yöntemi ile Bayesçi temel bileşen çözümlemesi yönteminin diğerlerine göre çok daha iyi sonuçlar verdiği görülmüştür. Gen sayısı düzeyinde çeşitli yüzdelik oranlarda eksik değer içeren veri kümelerinde Bayesçi temel bileşen çözümlemesi en iyi sonuçları verirken; örnek sayısı düzeyinde eksik değer içeren veri kümelerinde yerel en küçük kareler yöntemi en iyi sonuçları vermiştir. Yerel en küçük kareler ile Bayesçi temel bileşen çözümlemesi yöntemlerinin diğerlerine göre en önemli üstünlüğü veri kümesinin karmaşıklığından en az etkilenen yöntem olmalarıdır.**

Anahtar Kelimeler — mikrodizi; gen ifade verisi; eksik verilerin kontrolü ve düzeltilmesi; eksik değer kestirimi; en yakın k-komşu; Bayesçi temel bileşen çözümlemesi; yerel en küçük kareler; ortalama; ortanca.

***Abstract*—Control and correction process of missing values (imputation of MVs) is the first stage of the preprocessing of microarray datasets. This paper focuses on a comparison of most reliable and up to date estimation methods to control and correct the missing values. Imputation of MVs has a very high priority because of its impact on next pre-processing and post-processing stages of microarray data analysis namely, quality control, normalization, differential gene expression, classification, clustering, and pathway analysis, etc. Normalized root mean square error (NRMSE) value is used to evaluate the performances of most popular five methods (k-nearest neighbors, Bayesian principal component analysis, local least squares, mean and median). When NRMSE values of methods were compared, it has observed that local least squares (LLS) and Bayesian principal component analysis (BPCA) methods outperformed all other methods in all percentages of MVs (1%, 5%, 10%, and 20%). BPCA method has given the best results in all percentages of MVs over the number of probes or genes, whereas LLS method has given the best results in all percentages of MVs over the number of samples. The advantage of these two methods over others is that they are least affected by the complexity of the data set.**

Keywords — microarray; gene expression data; missing data imputation; missing value estimation; k-nearest neighbor; Bayesian principal component analysis; local least squares; mean; median.

I. GİRİŞ

DNA'nın keşfi ile birlikte, genetik kodun canlıların yaşamı üzerindeki etkileri en önemli araştırma alanlarından biri olmuştur. Gelişen teknoloji ile beraber bu alanda yapılan çalışmalarda da ciddi adımlar atılmıştır. Gen ifade verilerinin çözümlenmesi için geliştirilen birçok yöntemden biri olan mikrodizi teknolojisi, bir organizmaya ait tüm genlerin (genom) tek seferde çözümlenmesine olanak sağlamasıyla son yıllarda yaygın kullanıma sahip bir teknoloji olmuştur. Mikrodiziler, binlerce farklı DNA, protein, peptid, vs. parçalarının sentez edildiği veya yerleştirildiği binlerce noktadan (spot) oluşan çipler olarak

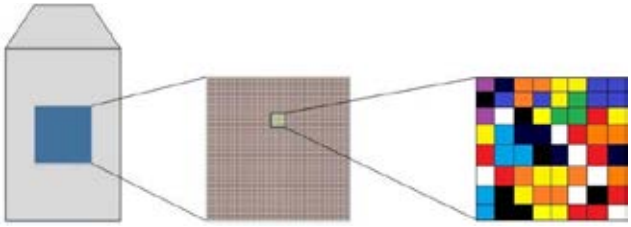
978-1-5090-2386-8/16/$31.00 ©2016 IEEE

da tanımlanabilir. Çip olarak ifade edilen bu katı yüzey, cam, plastik ya da silikondan yapılabilmektedir (Şekil 1'de solda yer alan görüntü). Çip yüzeyindeki her bir nokta "prob" olarak ifade edilmektedir. Çözümlemesi yapılacak hedef organizmaya ait gen dizisi verildiğinde baz eşleşmesi prensibi doğrultusunda hibridizasyon gerçekleşmektedir. Şekil 1'de ortada gösterilen görüntü tarama işlemi ile elde edilen görüntüdür. Şekil 1'de sağda yer alan büyütülmüş görüntü, (tamamı) binlerce özellikten (feature) yüzlercesini sunmakta olup şiddet değerlerine göre değişen renk ölçeğini göstermektedir. Bu renk ölçeğinde siyah renk, şiddetin sıfır olduğu bir diğer ifade ile hibridizasyonun gerçekleşmediği anlamını taşımaktadır. Affymetrix platformunda bu renk ölçeği en düşük şiddetten en yüksek şiddet değerine göre koyu mavi renkten başlayıp kırmızıya sonra beyaza kadar değişmektedir. Tarama işleminin ardından, elde edilen görüntü verisi çeşitli istatistiki ve/veya veri madenciliği yöntemleri ile çözümlenmektedir [1]. Mikrodizi teknolojisi, yirmi yıllık geçmişine rağmen hala eksik değer (missing value) içerebilmektedir [2-5]. Fabrikasyon hataları, zayıf hibridizasyon, yetersiz çözünürlük veya çip (chip) üzerindeki kirler eksik verilerin oluşum sebeplerinden bazılarıdır. Eksik değerlerle ilgili yapılan en basit işlem, eksik değer içeren probları gözlem verisinden çıkarmak olmuştur. Fakat bu yöntem (1) çok az sayıda eksik değer içeren prob sayısı söz konusu olduğunda veya (2) tüm örneklerin çözümlemesi yapılırken, yorumlama süresince ciddi bir ön yüklemeye (bias) sebep olmayacaksa yararlı olabilir [6-8].

Eksik verilerin kontrolü ve düzeltilmesi, bu eksik değerlerin yerine kestirilen yeni değerlerin yerleştirilmesini amaçlayan bir dizi yöntemi içermektedir. Çoğu durumda, bir veri kümesinin nitelikleri birbirinden bağımsız değildir. Bu yüzden, nitelikler arasındaki ilişkilerin tanımlanması ile eksik değerler belirlenebilir. Eksik veri problemini çözmenin basit bir çözümü mikrodizi deneylerini tekrarlamaktır. Ancak bu çözüm çok maliyetli ve verimsizdir. Bir başka çözüm de, bir veya daha fazla eksik değer içeren probları kaldırmaktır ki bu işlemin önemli bilgilerin de kaybolmasına sebep olacağı açıkça anlaşılmaktadır. Sonuç olarak, eksik değerlerin doğru şekilde düzeltilmesi için daha gelişmiş algoritmaların geliştirilmesi zorunludur.

Önceden yapılan çalışmalar incelendiğinde birçok eksik değer düzeltme yöntemlerinin sunulduğu görülmüştür [4, 5]. İlk yaklaşımlar eksik değerlerin yerine sıfır veya satır/sütun ortalama değerlerinin (aynı zamanda ortalama ve ortanca değer yöntemleri olarak da ifade edilmektedir.) yerleştirilmesi üzerinedir. Bu en basit yöntemler veri kümesinde korelasyon ilişkisini hesaba katmamaktadır. Bu durum, kestirim doğruluğunun azalmasına sebep olmaktadır [8, 9].

Eksik değer kestirimi için mikrodizi verilerinin çözümlenmesinde başarımı arttıran ve yüksek doğruluğa sahip olan ileri algoritmalara ihtiyaç duyulmaktadır. Son zamanlarda, eksik değerlerin kestirim başarımını belirgin farkla iyileştiren bu gelişmiş algoritmalara çeşitli türlerde biyolojik verileri içeren çalışmalarda rastlanmıştır. Ancak, önceden yapılan çalışmalarda da sunulduğu gibi, tüm veri kümesi türleri için iyi çalışan bir algoritma henüz geliştirilememiştir. Sonuç olarak, en iyi algoritma bulunamaz, fakat her veri kümesi için uygun değer (optimal) algoritmaları bulunabilir [5, 11]. Geçtiğimiz son 10 yıl içinde, eksik değer düzeltme uygun değer algoritmaları geliştirmek için çeşitli çalışmalar yapılmıştır [3-11].



Şekil 1. Kartuş içinde paketlenmiş temsili bir gen çipi (solda), prob kümesi (ortada) ve probların flüoresan görüntüsü (sağda).

Bu çalışmada amacımız, mikrodizi veri kümesinde eksik değer kontrolünü ve düzeltilmesini sağlamak için en yakın k-komşu (KNN), Bayesçi temel bileşen çözümlemesi (BPCA), yerel en küçük kareler (LLS), ortalama (MEAN) ve ortanca (MEDIAN) yöntemlerinin başarımlarını karşılaştırmaktır. Gerçek veri kümesinden %1, %5, %10 ve %20 oranlarında rasgele şekilde eksik değer içeren yapay veri kümeleri oluşturulmuştur. 5 yöntemin her biri ile eksik değerlerin yerine üretilen yeni değerler yerleştirilerek veri kümesinde eksik değer problemi giderilmiştir. Bu yeni veri kümesi, eksik değerli yapay veri kümesi ve gerçek veri kümesi kullanılarak normalize edilmiş karekök ortalama hata (normalized root mean square error, NRMSE) değerleri hesaplanarak başarımları karşılaştırılmıştır. Çözümleme ve başarım sonuçları ile ilgili ayrıntılı bilgi gelecek bölümlerde sunulmuştur. Tüm çözümlemeler R ortamında programlama ile gerçekleştirilmiştir.

II. YÖNTEM

A. Gerçek ve Yapay Veri Kümeleri

Bu çalışmada kullanılan gerçek veri kümesi 52 adet örnek (sütun) ve 500 adet genin (satır) prob ifadelerini (probe or spot intensity) içermektedir. Gerçek veri kümesi eksik değer içermemektedir. Yapay veri kümeleri ise sırasıyla %1, %5, %10 ve %20 oranlarında rasgele şekilde değerler eksiltilerek oluşturulmuştur.

B. Kestirim Yöntemleri

Bu bölümde en popüler kestirim yöntemlerinin çalışmamızda nasıl kullanıldıkları hakkında teorik ve pratik bilgi sunulacaktır.

KNN algoritması uygulanırken öncelikle komşu sayısı *k* değeri belirlenir. Diğer belirli gen ifade verilerinden hedefteki eksik değere olan Öklid uzaklıkları hesaplanır. Hesaplanan bu uzaklıklar sıralanır ve en küçük uzaklığa bağlı olarak en yakın komşular bulunur. En yakın komşu kategorileri toplanır ve en uygun komşu kategorisi seçilir. *n*

boyutlu uzayda Öklid uzaklığının hesaplanması Eşitlik 1 ile sağlanır.

$$u(k,l) = u(l,k) = \sqrt{(k_1 - l_1)^2 + (k_2 - l_2)^2 + \cdots + (k_n - l_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (k_i - l_i)^2} \quad (1)$$

Burada $i=1,2, \ldots n$ için $k=(k_1, k_2, \ldots, k_n)$ ve $l=(l_1, l_2, \ldots, l_n)$ Kartezyen koordinatlarına sahip k ve l değerlerinin arasındaki uzaklığın hesaplanmasında Pisagor Teoremi esas alınır. KNN yönteminin avantajı, artan eksik değer oranına rağmen başarımında daha az bozulma göstermesidir.

BPCA yöntemi, temel bileşen çözümlemesi (Principal Component Analysis, PCA) için yinelemeli iyileştirme algoritması olan beklenti – ençoklama (Expectation – Maximization, EM) algoritması ile Bayesçi modelini birleştirir. Standart PCA'da eğitim kümesinden uzak, temel alt uzaya yakın olan veri aynı geri çatma hatasına (reconstruction error) sahip olabilir. BPCA geri çatma hata değerlerine göre benzerlik (likelihood) hesaplar. BPCA yönteminin avantajı, kendisinden önce geliştirilen yöntemlerden (özellikle de KNN tabanlı yöntemden) belirgin farkla daha iyi sonuçlar vermesidir. Sebebi ise, diğer yöntemlerin belirlenmesi oldukça zor olan model parametrelere ihtiyaç duyması; BPCA yönteminin parametrelere ihtiyaç duymamasıdır. BPCA yöntemi özelikle mikrodizi çalışmalarında eksik değer düzeltme işlemi için geliştirilmiştir [12].

LLS yöntemi, eksik değerli genin en yakın k-komşularının doğrusal bir kombinasyonuna dayanır. Değişkenler arasındaki uzaklık Pearson, Spearman veya Kendall korelasyon sabitlerinin mutlak değeri olarak tanımlanır (Pearson korelasyon sabiti bu çalışmada en iyi sonuç veren sabit olmuştur.). LLS yöntemi ile BPCA yönteminin en büyük farkı, LLS yönteminin yerel benzerlik yapısına dayanan bir optimizasyon süreci olması; BPCA yönteminin ise temel bileşenlere dayanan bir optimizasyon süreci olmasıdır.

Ortalama ve ortanca yöntemleri eksik değerli genin tüm örneklerdeki değerlerinin (satır) ortalaması veya ortanca değerinin hesaplanmasına dayanır.

C. Başarım Ölçütü

Algoritma ve yöntemlerin başarımlarını değerlendirmek için NRMSE değerleri kullanılmıştır. NRMSE, gerçek ve yapay veri kümelerindeki değerler arasındaki benzerlikleri değerlendiren ve çok popüler kullanıma sahip olan bir indistir. Eşitlik 2'de bu indisin formülü verilmiştir.

$$NRMSE = \sqrt{\frac{\text{ortalama}[(y_t - y_g)^2]}{\text{varyans}[y_g]}} \quad (2)$$

Burada, y_t indisi eksik değerin yerine yerleştirilen yeni kestirim değerini; y_g indisi gerçek değeri ifade etmektedir.

III. SONUÇLAR VE TARTIŞMA

Gerçek veri kümesinden %1, %5, %10 ve %20 oranlarında rasgele şekilde eksik değer içeren yapay veri kümeleri oluşturulmuştur. 5 yöntemin her biri ile eksik değerlerin yerine üretilen yeni değerler yerleştirilerek veri kümesinde eksik değer problemi giderilmiştir. Bu yeni veri kümesi, eksik değerli yapay veri kümesi ve gerçek veri kümesi kullanılarak hesaplanan NRMSE değerleri yöntemlerin başarımlarını karşılaştırmada önemli bir indis olmuştur.

Bu çalışmada, KNN yöntemi için tüm yapay veri kümelerinde en iyi k değeri 10 olarak bulunmuştur ve bu değer için yöntemin kestirim yapması sağlanmıştır. LLS yöntemi için elde edilen sonuçlardan sadece %20 eksik değer içeren veri kümesi değerlendirilirse, LLS yöntemi uygulanırken korelasyon sabitleri olarak Pearson, Kendall ve Spearman sabitleri kullanıldığında elde edilen NRMSE değerleri sırasıyla 0.178, 0.179 ve 0.179'dur. Sonuç olarak, LLS yönteminde korelasyon sabiti olarak Spearman, Kendall ve Pearson sabitleri arasından en iyi sonuç veren Pearson sabiti kullanılmıştır.

Yöntemlerin örnek sayısı ve gen sayısı üzerinden eksik değer oranlarına karşı dayanıklılıklarını karşılaştırmak için önce örnek sayısı düzeyinde %50 oranı sabit tutularak; gen sayısı düzeyinde sırasıyla %2, %10, %21 ve %42 oranlarında değerler eksiltilmiş ve böylelikle %1, %5, %10 ve %20 oranlarında eksik değer içeren yapay veri kümeleri oluşturulmuştur. Çizelge 1'de 4 farklı yapay veri kümesi için kestirim yöntemlerinin NRMSE değerleri yer almaktadır. Bu değerler incelendiğinde hemen hemen tüm yapay veri kümelerinde BPCA yönteminin diğerlerine göre çok daha iyi sonuç verdiği belirgin farkla görülmektedir. LLS yönteminin de BPCA yöntemi kadar etkili sonuçlar verdiği yine çizelgedeki sonuçlara bakıldığında anlaşılmaktadır.

ÇİZELGE 1. YÖNTEMLERIN NRMSE DEĞERLERI (ÖRNEK SAYISI DÜZEYINDE %50 ORANINDA SABIT; GEN SAYISI DÜZEYINDE DEĞIŞKEN EKSIK DEĞER)

Eksik Değer (%) / Kestirim Yöntemleri (NRMSE Değerleri)	Kestirim Yöntemleri				
	KNN	BPCA	LLS	MEAN	MEDIAN
1	0.151	**0.137**	0.143	1.023	1.063
5	0.181	**0.143**	0.144	1.029	1.008
10	0.209	**0.187**	0.200	1.000	1.019
20	0.235	0.188	**0.178**	1.007	1.027

İkinci olarak, gen sayısı düzeyinde %50 eksik değer oranı sabit tutularak; örnek sayısı düzeyinde sırasıyla %2, %12, %21 ve %40 oranlarında değerler eksiltilmiş ve böylelikle %1, %5, %10 ve %20 oranlarında eksik değer içeren yapay veri kümeleri oluşturulmuştur. Çizelge 2'de 4 farklı yapay veri kümesi için kestirim yöntemlerinin NRMSE değerleri yer almaktadır.

ÇİZELGE 2. YÖNTEMLERIN NRMSE DEĞERLERI (GEN SAYISI DÜZEYINDE %50 ORANINDA SABIT; ÖRNEK SAYISI DÜZEYINDE DEĞIŞKEN EKSIK DEĞER)

Eksik Değer (%) / Kestirim Yöntemleri (NRMSE Değerleri)	Kestirim Yöntemleri				
	KNN	BPCA	LLS	MEAN	MEDIAN
1	0.155	0.126	**0.116**	1.003	0.998
5	0.232	0.185	**0.162**	1.004	1.019
10	0.234	0.193	**0.169**	1.004	1.024
20	0.231	0.191	**0.176**	1.005	1.025

Bu değerler incelendiğinde tüm yapay veri kümelerinde LLS yönteminin diğerlerine göre çok daha iyi sonuç verdiği belirgin farkla görülmektedir. BPCA yönteminin de LLS yöntemi kadar etkili sonuçlar verdiği yine çizelgedeki sonuçlara bakıldığında anlaşılmaktadır. KNN yöntemi ile de elde edilen veri kümelerinin NRMSE değerleri LLS ve BPCA yöntemlerine oldukça yakın elde edilmiştir.

Sonuç olarak ortalama ve ortanca yöntemleri gen ifade verilerinin varyasyonunu değiştirmediğinden eksik değerlerin yerine sıfır koyma yöntemine göre tercih edilmesi gereken yöntemlerdir. Bu yöntemlerin yerine son 10 yıl içerisinde geliştirilen ileri algoritmaların tercih edilmesi gerektiğini belirgin farkla sunan çalışmamız, araştırmacılara bir rehber çalışma niteliğindedir. Bilindiği üzere, mikrodizi veri kümelerinde gen sayısı örnek sayısından çok daha fazladır. Bu durum dikkate alınırsa Çizelge 1'de gen sayısına göre önemli oranlarda eksik değer içeren veri kümelerinde BPCA yönteminin oldukça iyi sonuçlar ürettiği; LLS yönteminin de ona yakın derecede önemli sonuçlar verdiği gözlenmektedir. Çizelge 2'de ise, gen sayısı üzerinden belirli bir oranda eksik değer var iken, örnek sayısı üzerinden çeşitli oranlarda eksik değer içeren veri kümelerinde LLS yönteminin belirgin şekilde en iyi sonuçları verdiği; yine BPCA yönteminin önemsenecek derecede iyi sonuçları verdiği görülmektedir. LLS ve BPCA yöntemleri diğerlerine göre çok iyi sonuçlar verdiğinden daha farklı mikrodizi veri kümeleri için de öncelikli olarak tercih edilebilirler. KNN yöntemi ile de üretilen veri kümeleri kullanılabilecek düzeyde olup, NRMSE değerleri LLS ve BPCA yöntemlerine oldukça yakın elde edilmiştir.

KAYNAKÇA

[1] Özkan, Y., Selçukcan Erol, Ç. "Biyoenformatik, DNA, Mikrodizi,Veri Madenciliği", Papatya Yayıncılık, 2015.

[2] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. "Missing value estimation methods for DNA microarrays", Bioinformatics, 17(6): 520–5, 2001.

[3] Celton, M., Malpertuy, A., Lelandais, G., de Brevern, A. "Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments", BMC Genomics, 11(1): 15, 2010.

[4] Oh, S., Kang, D.D., Brock, G.N., Tseng, G.C. "Biological impact of missing-value imputation on downstream analyses of gene expression profiles". Bioinformatics, 27(1): 78–86, 2011.

[5] Chiu, C.C., Chan, S.Y., Wang, C.C., Wu, W.S. "Missing value imputation for microarray data: a comprehensive comparison study and a web tool", BMC Syst Biol. 7(S-6): 12, 2013.

[6] Little, R.J.A., Rubin, D.B. "Statistical analysis with missing data", NJ: Wiley, 2002.

[7] Luengo, J., García, S, Herrera, F. "On the choice of the best imputation methods for missing values considering three groups of classification methods". Knowl Inf Syst. 2012; 32(1):77–108.

[8] de Souto, M.C.P., Jaskowiak, A.P., Costa, I.G. "Impact of missing data imputation methods on gene expression clustering and classification". BMC Bioinformatics 16:64, 2015.

[9] Alizadeh, A.A. et.al. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling". Nature, 403(6769):503–11, 2000.

[10] Tuikkala. J., Elo, L.E., Nevalainen, O.S., Aittokallio, T. "Missing value imputation improves clustering and interpretation of gene expression microarray data". BMC Bioinformatics 9:202, 2008.

[11] Brock, et. al. "Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes", BMC Bioinformatics 9:12, 2008.

[12] Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K., Ishii, S. "A Bayesian missing value estimation method for gene expression profile data", BMC Bioinformatics 9:16, 2088–2096, 2003.