YBS Ansiklopedi

www.YBSAnsiklopedi.com

Cilt 4, Sayı 3, Eylül 2017

Eksik Verilerin Tamamlanması (Imputation)

Şadi Evren ŞEKER[1], Eda Eşmekaya[2]

1. *İstanbul Şehir Üniversitesi, Yönetim Bilişim Sistemleri Bölümü*
2. *Süleyman Demirel Üniversitesi, Bilgisayar Mühendisliği Bölümü*

Özet

Bu yazının amacı, bilgisayar bilimlerinde özellikle veri madenciliği (data mining) konularında eksik verilerle karşılaşılması halinde bir çözüm olarak bu eksik verilerin töhmet edilmesi, eksik verilerin tamamlanması (yerine uygun verilerin üretilmesi, imputation) yöntemini açıklamaktır. ETL süreçlerinin bir parçası olarak veri kaynaklarından çıkarılan (extract) eksik verilerin ilerideki problemlere sebep olmadan çözülmesi sürecine töhmet ismi verilmektedir. Bu yazı kapsamında, literatürde sık kullanılan ve endüstrideki veri ambarı veya veri bilimi dönüşümlerinde de sıkça başvurulan 9 farklı töhmet yöntemi örnekler üzerinden açıklanacaktır.

Anahtar Kavramlar:*veri ön işleme ,veri tabanı , veri madenciliği, veri bilimi*

Abstract

The aim of this article is to explain the imputation techniques which is widely used in data mining studies. It is also one of the mandatory parts for most of the ETL processes . This paper covers the most widely used methods in the industry for data warehouse or data science purposes. 9 of these methods are mentioned in the paper.

Keywords: *data preprocessing, data bases, data mining, data science*

1. Giriş

Töhmet, sözlükte olmayan bir şeyin yüklenmesi anlamındadır. Örneğin olmayan bir suçun birisine yüklenmesine töhmet altında bırakmak denilebilir. Bu anlamda bir veri kümesi (data set) üzerinde çalışılırken bazı sebeplerden dolayı verilerin eksik olması halinde bu verilerin uygun başka sayılarla tamamlanması sağlanabilir [1].

Genelde verinin hatalı okunması, veri kaynağında yaşanan bozulma gibi sorunlar veya bazı verilere erişim zorluğu eksik verilere sebep olabilmektedir.

Bu verilerin eksik olması durumu genelde sorunlara sebep olur. Örneğin çoğu hazır istatistik paketleri (SAS, SPSS, Weka veya r-project gibi) bu tip durumlarda sorunlar yaşamaktadır. Gerçi bu paketlerin ücretli ve gelişmiş olanlarının çoğunda (SAS veya SPSS gibi) töhmet modülleri (imputation) bulunmakta ve bu işi otomatik olarak

yapabilmektedirler ancak bu yazı kapsamında gerek bu modüllerin nasıl çalıştığını anlamaya çalışacağız gerekse bu işlemi elle yapmak istediğimizde nasıl müdahale etmemiz gerektiğini açıklamaya çalışacağız.

Töhmet yöntemleri duruma ve beklentilerimize göre çeşitlilik arz eder ve halen üzerinde çalışılmaktadır. Bunlardan çok bilinen bazılarını saymamız gerekirse:

- Sıcak deste (hot deck)
- Soğuk deste (cold deck)
- Liste boyunca silme (listwise deletion)
- Eşlerin silinmesi (pairwise deletion)
- Ortalama töhmet (mean imputation)
- İlkelleme töhmedi (regression imputation)
- Son gözlemin taşınması (last observation carried forward)
- Olasılıksal töhmet (stochastic imputation)
- Çoklu Töhmet (Multiple imputation)

Yukarıdaki bu kavramları kısaca açıklamaya çalışalım. Öncelikle sıcak deste ve soğuk deste (hot deck , cold deck) algoritmaları için ne yazık ki tam bir mutabakat olmadığını ve çeşitli versiyonları bulunduğunu belirtmek gerekir (bkz. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3130338/) Bu yüzden bu iki algoritma bu yazı kapsamında ele alınmayacaktır, ancak hot deck (sıcak deste) için kısaca rast gele bir verinin seçilerek eksik olan veri yerine töhmet edildiğini söylemek yeterli olacaktır [2][3]. Bu rast gele seçim ise oldukça tartışmalı bir konu. Diğer yöntemler ise aşağıdaki şekilde açıklanabilir.

2. Liste boyunca silme (listwise deletion)

Örnek bir veri tabanı tablosu olarak Tablo 1'de gösterilen tabloyu ele alalım.

Tablo 1. Örnek Veri Tabanı Tablosu

İsim	Yaş	Kısım
Şadi Evren ŞEKER	33	Bilgisayar
Ali Demir	23	Muhasebe
Cem Yıldız	26	Bilgisayar
Ahmet Yılmaz	33	——
—-	—	Muhasebe
Veli Demir	43	Yönetim

Tablo 1'de sondan ikinci ve üçüncü satırlarda kayıp bilgi bulunmaktadır. Bu durumda, örneğin çalışanların yaş ortalamalarının alınması istendiğinde hata ile karşılaşılacaktır.

Liste boyunca silme yaklaşımında (listwise deletion) bu kayıp veri içeren satırların tamamı tablodan temizlenir ve veri tabanındaki tablo, Tablo 2'de gösterilen hale getirilir.

Tablo 2. Liste Boyunca Silme işlemi sonrası

İsim	Yaş	Kısım
Şadi Evren ŞEKER	33	Bilgisayar
Ali Demir	23	Muhasebe
Cem Yıldız	26	Bilgisayar
Veli Demir	43	Yönetim

Tablo 2'deki haliyle, veri tablosunda herhangi bir eksik veri bulunmamaktadır ve verinin işlenmesi sırasında eksik verilerden kaynaklanan herhangi bir sorunla karşılaşılmaması beklenir. Liste boyunca silme problemi, veri tabanındaki faydalı olabilecek ve veri bilimi açısından işlenirken sonuçları olumlu yönde etkileyecek diğer verilerin kaybolmasına sebep olduğu için dezavantajları olan bir yöntemdir. Örneğin herhangi bir kolonda çok sayıda eksik veri bulunması durumunda (ki bu eksik oranı bazı durumlarda %90'lara kadar varabilmektedir), veri tabanının önemli bir kısmını kaybetmekle sonuçlanabilir [4].

3. Eşlerin silinmesi (Pairwise deletion)

Bu yöntemde bütün tablonun temizlenmesi yerine gerekli olan işlem sırasındaki eksik veriler temizlenir. Örneğin bir önceki tabloya geri dönecek olursak ve yapmak istediğimiz işlem, yaş ortalamasını hesaplamak ise:

Tablo 3. Eşlerin Silinmesi işlemi öncesi

İsim	Yaş	Kısım
Şadi Evren ŞEKER	33	Bilgisayar
Ali Demir	23	Muhasebe
Cem Yıldız	26	Bilgisayar
Ahmet Yılmaz	33	——
—-	—	Muhasebe
Veli Demir	43	Yönetim

Bu durumda, bu soruya özel olarak sadece sondan ikinci satırın silinmesi yeterlidir:

Tablo 4. Eşlerin Silinmesi işlemi sonrası

İsim	Yaş	Kısım
Şadi Evren ŞEKER	33	Bilgisayar
Ali Demir	23	Muhasebe
Cem Yıldız	26	Bilgisayar
Ahmet Yılmaz	33	——
Veli Demir	43	Yönetim

Artık istenen işlem yani yaşların ortalaması çalıştırılabilir.

Bu yöntemde farklı bir işlem yapılmak istendiğinde, örneğin çalışanların kısımlara göre dağılım grafiği istendiğinde ise işlem yapacağımız tablo aşağıdaki şekilde olacaktır:

Tablo 5. Farklı amaçla Eşlerin Silinmesi işlemi Sonrası

İsim	Yaş	Kısım
Şadi Evren ŞEKER	33	Bilgisayar
Ali Demir	23	Muhasebe
Cem Yıldız	26	Bilgisayar
—-	—	Muhasebe
Veli Demir	43	Yönetim

Görüldüğü üzere sadece o işlem için problem çıkaran satır silinmiş diğer satırlar farklı kolonlarında eksik veri bulunmasına rağmen saklanmıştır.

Bu yöntemin menfi yanı, her işlem için farklı veri kümesi üretiliyor olması ve hatta her işlem sonucunda farklı sayıda veri ele alınıyor olmasıdır. İşlemler sonucunda verilerin karşılaştırılma güçlüğü ortaya çıkabilir.

4. Ortalama Töhmet (Mean imputation)

Bu yöntemde, bir eksik verinin töhmeti sırasında ortalama değer hesaplanır. Örneğin yine aynı tablo üzerinden töhmet yaklaşımını izah edelim:

Tablo 6. Ortalama Töhmet Öncesi

İsim	Yaş	Kısım
Şadi Evren ŞEKER	33	Bilgisayar
Ali Demir	23	Muhasebe
Cem Yıldız	26	Bilgisayar
Ahmet Yılmaz	33	——
Veli Demir	43	Yönetim

Şayet bu tablodaki ortalama yaş için eksik olan satırların töhmeti söz konusuysa bu durumda satırların silinmesi yerine veri üretilir. Örneğimizdeki yaşların ortalaması:

$$ortalama\ yaş = \frac{33+23+26+33+43}{5} = 31.6$$

olarak bulunur. Yaşın ondalıklı sayı olması mümkün olmadığı için 32 olarak yuvarlanarak kabul edilebilir ve veri kümemiz aşağıdaki şeklini alır:

Tablo 7. Ortalama Töhmet Sonrası

İsim	Yaş	Kısım
Şadi Evren ŞEKER	33	Bilgisayar
Ali Demir	23	Muhasebe
Cem Yıldız	26	Bilgisayar
Ahmet Yılmaz	33	——
Veli Demir	43	Yönetim

Bu işlemin ardından artık istenen veri madenciliği yöntemleri uygulanabilir.

Bu yöntemin dezavantajı ise, çok büyük veriler de uygulanma zorluğudur (örneğin güncel problemlerin artık haritalama-indirgeme (map reduce) ortamlarında işlendiği düşünülürse, böyle bir ortamda kullanılamaz). Ayrıca işlenmesi için bütün verinin hafızaya yüklenip hesaplama yapılması zorluğu da bulunmaktadır.

Son olarak veri eksikliğinin çok fazla olduğu durumlarda üretilen verilerin sağlığı problem çıkarmaktadır. Ayrıca genelde nümerik veri tipleri için uygun bir yöntem olmakla birlikte nominal veri tipleri için ortalama hesaplamak mümkün olmadığından dolayı bir çözüm sunamamaktadır.

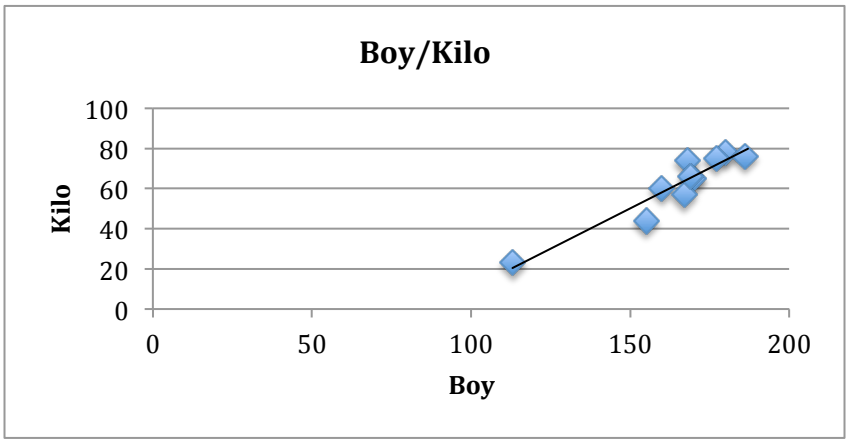
5. İlkelleme Töhmeti (Regression Imputation):

Bu yöntemde, mevcut veriler üzerinden fonksiyonel bir ilkelleme (regerssion) yapılır (örneğin doğrusal ilkelleme (linear regression)) ve ardından ilkellenen fonksiyondan üretim yapılarak eksik veri doldurulur. Ancak ilkelleme fonksiyonlarının çalışabilmesi için genelde en az 2 nümerik alana ihtiyaç duyulur.

Tablo 8. Boy / Kilo ve Cinsiyet tablosu

boy	kilo	cinsiyet
180	78	erkek
177	75	erkek
170	65	kadın
168	74	erkek
186	76	kadın
187	——	erkek
167	57	kadın
160	60	kadın
113	23	erkek
169	66	erkek
155	44	kadın

Örneğin Tablo 8'de boy ve kilo olmak üzere iki nümerik veri bulunmaktadır. Eksik olan kilo değerini bulabilmek için diğer satırları kullanarak doğrusal ilkelleme yapılabilir.



Şekil 1: Veri Tablosunun Doğrusal İlkellemesi

Bu ilkelleme sonucunda eksik olan 187 boyun karşılığı olan kilo, üretilen doğru üzerinde aranarak bulunur. Örneğimizde bu değer 81 olarak okunacaktır ve tablonun eksik verisinin tamamlanmış hali aşağıdaki şekilde olacaktır.

Tablo 9. Doğrusal İlkelleme Sonucu

boy	kilo	cinsiyet
180	78	erkek
177	75	erkek
170	65	kadın
168	74	erkek

186	76	kadın
187	81	erkek
167	57	kadın
160	60	kadın
113	23	erkek
169	66	erkek
155	44	kadın

6. Son Gözlemin Taşınması (Last Observation Carried Forward):

Bu yöntemde, veri kümesindeki eksik veriler, kendilerinden bir önceki veri kopyalanmak marifetiyle töhmet edilirler. Örneğimize geri dönecek olursak:

Tablo 10. Son Gözlemin Taşınması Öncesi

İsim	Yaş	Kısım
Şadi Evren ŞEKER	33	Bilgisayar
Ali Demir	23	Muhasebe
Cem Yıldız	26	Bilgisayar
Ahmet Yılmaz	33	——
—-	—	Muhasebe
Veli Demir	43	Yönetim

Veri kümesindeki verilerin son gözlemin taşınması yöntemiyle töhmet edilmiş hali aşağıdaki şekildedir:

Tablo 11. Son Gözlemin Taşınması Sonrası

İsim	Yaş	Kısım
Şadi Evren ŞEKER	33	Bilgisayar
Ali Demir	23	Muhasebe
Cem Yıldız	26	Bilgisayar
Ahmet Yılmaz	33	Bilgisayar
Ahmet Yılmaz	33	Muhasebe
Veli Demir	43	Yönetim

Artık bu veri kümesi üzerinde veri madenciliği yöntemleri uygulanabilir.

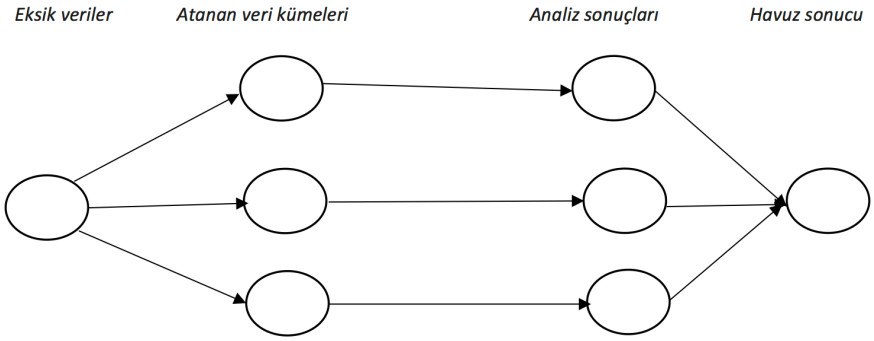
Bu yöntemin dezavantajı ise bazı marjinal noktaların çoğaltılması olarak görülebilir. Örneğin Yaş sütununda marjinal olarak 70 yaşında bir çalışan varsa bu kişinin kopyalanması, veri madenciliği sonuçlarını menfi etkileyebilir.

Son iki yöntemimiz olan istatistiksel töhmet (stochastic imputation) ve çoklu töhmet (multiple imputation) için şunları söylememiz yeterli olacaktır.

Öncelikle mevcut veri kümesi üzerinden bir istatistiksel dağılım elde edilir (örneğin ilkelleme (regression) burada kullanılabilir) ardından bu dağılım sayesinde eksik verileri dolduran bir bağlantı kullanılır. Çoklu töhmette ise bu işlem birden fazla kere yapılarak her bir veri kümesi daha sonra kullanılmak üzere saklanır. Ayrıca her veri kümesinin hata miktarı hesaplanır. Ardından bu veri kümelerinin ortalama değerleri alınarak nihai veri kümesi elde

edilir. Buradaki tek ayrıntı standart hatanın (standard error) hesaplanması sırasında iki veri kümesinin birleştirilmesi için iki standart hata miktarı toplanıp karekökleri alınır.

Kısacası çoklu töhmet, daha iyi sonuç elde etmek için istatistiksel töhmetin birden fazla kere çalıştırılması olarak düşünülebilir [5].



Şekil 2: Çoklu Töhmet Akış Şeması

Tüm çoklu atama metotları 3 adımda gerçekleşir:

- Töhmet: Tekli atama (single imputation) benzerdir, eksik veriler atanır. Fakat atanılan veriler bir kez yerine "m" zamanda dağıtılır. Bu adımın sonunda "m" veri kümeleri tamamlanmalıdır.
- Analiz: "m" veri kümesinin her biri analiz edilir. Bu adımın sonunda m analiz olmalıdır.
- Havuzlama (pooling) :"m" sonuçları, verinin ortalama, varyans ve güvenirlik aralığı hesaplanarak bir sonuca bağlanılır.

7. İleri Töhmet Yöntemleri

Yukarıdaki töhmet yöntemlerine ilave olarak, herhangi bir makine öğrenmesi algoritması da töhmet amacıyla kullanılabilir. Örneğin en yakın k komşu algoritması kullanılarak eksik verilerin hesaplanması mümkündür. Şayet k değeri 1 ise, algoritma basitçe, elindeki verilere bakarak en yakın komşuyu bulacak ve eksik olan veriyi bu komşudaki veri ile dolduracaktır. K, parametresinin daha yüksek değerler alması durumunda ise ortalama almak veya verinin özelliklerine göre farklı çarpanlarla ortalamaya etki etmek gibi yöntemler geliştirilebilir.

Makine öğrenmesi algoritmalarından, tahmin (prediction) için kullanılan herhangi bir algoritma bu anlamda eksik verilerin tahmini içinde başarılı bir şekilde çalışacaktır. Hatta, yapay sinir ağları ve derin öğrenme yaklaşımları da töhmet problemlerine başarılı şekilde uyarlanabilmektedir.

Kaynakça

[1] Allison, Paul D. "Missing data: Quantitative applications in the social sciences." British Journal of Mathematical and Statistical Psychology 55.1 (2002): 193-196.

[2] Rao, Jon NK, and Jun Shao. "Jackknife variance estimation with survey data under hot deck imputation." Biometrika 79.4 (1992): 811-822.

[3] Kalton, Graham. "Compensating for missing survey data." (1983).

[4] Seker, Sadi Evren, and Enes Eryarsoy. "Generating Digital Reputation Index: A Case Study." Procedia-Social and Behavioral Sciences 195 (2015): 1074-1080.

[5] Sterne, Jonathan AC, et al. "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls." Bmj 338 (2009): b2393.