

Veri Madenciliğinde Kayıp Veriler İçin Kullanılan Yöntemlerin Karşılaştırılması

Evren Sezgin¹, Yüksel Çelik²

¹Akdeniz Üniversitesi, Enformatik Bölüm Başkanlığı, Antalya.

²Karamanoğlu Mehmetbey Üniversitesi, Karaman MYO, Karaman.
esezgin@akdeniz.edu.tr vcelik@kmu.edu.tr

Özet: Araştırmalarda iki boyutlu veri setleri oluşturulurken veri değerleri çok önem arz etmektedir. Bazen bu veri setlerinde eksik değerler olması araştırmayı oldukça etkileyebilir. Bu sorundan doğabilecek hata toleransını azaltabilmek için bir çok yöntem geliştirilmiştir. En basit olarak kayıp veri olan kayıtları yok saymak bulunabilecek çözümlerden birisidir. Ancak kayıp verilerin çok olduğu bir sette bu yöntem hata oranını oldukça yükseltir. Bu yöntemin yerine kayıp veriyi; Regresyon ile belirleme, Hot/Cold Deck ile Belirleme, Beklenti Maksimizasyonu, Son Gözlemi İleri Taşıma, Çoklu Atama, Karar Ağacı, Naive Bayes gibi yöntemler kullanılabilir. Bu çalışmada bu yöntemlerin Avantaj ve Dezavantajları karşılaştırılmış, veri setleri üzerinde denemeler yapılarak, yöntemlerin verimliliği ölçülmüştür.

Anahtar Kelimeler: Veri madenciliği, Kayıp Veriler, Algoritmalar.

Comparison of Data Used For Loss Of Data Mining Methods

Abstract: Research data values when creating a two-dimensional data sets is very important. Sometimes this can greatly affect the research data sets have missing values. To reduce this problem, a lot of fault-tolerance methods have been developed that may arise. Most simply ignore the missing data record, which is one of the solutions can be found. However, there are a lot of lost data error rate of this method is quite raises a set. Instead of this method is that the data loss, the regression and determination, Hot / Cold Deck and Identification, Expectation Maximization, Last Observation Onward Transfer, Multiple Assignment, such as the decision tree methods can be used. Advantages and Disadvantages of these methods are compared in this study, performed experiments on data sets, methods, efficiency is measured.

Keywords: Data mining, Lost Data, Algorithms.

1.Giriş

Kayıp veri terminolojisi ilk kez Little ve Rubin tarafından kullanılmıştır [1].Bu çalışmada kayıp verileri - varsayımsal açıdan -oluşum nedenlerine göre 3 ana sınıfta değerlendirmişlerdir [1].

1. Tümüyle Raslantısal Kayıp (Missing Completely Random, MCAR): Verileri setlerini oluştururken tamamen istek dışı oluşan veri kayıplarıdır. Örneğin bir anket çalışmasında soruyu görmeyip cevaplamama veya verilerden bazılarının kaybolması gibi.

2. Raslantısal Kayıp (Missing at Random, MAR): Bir anket çalışmasında ; , örneklem bireylerini oluşturan grubun sorulan sorulara bilerek atlaması veya yanlış cevaplar vermesidir.

3. Raslantısal Olmayan Kayıp/ Gözardı Edilemez Kayıp (Missing not at Random, MNAR/ Non-ignorable Missing, NIM): Anketteki bir sorunun yanlış sorulmasından dolayı doğru cevabın çözülemediği sorular bu varsayıma örnektir.

2.Yöntemler

Gözlemlenebilen/Ölçülebilen bütün sistemler de belirli oranlarda - az ya da çok - kayıp veri varlığı kaçınılmazdır. Veri kaybı belirlenen ya da

belirlenemeyen pek çok nedenden kaynaklanabilir. Doğru ve güvenilir analizler için veri kümesinin eksiksiz olması oldukça önemlidir.

Kayıp verilerin değerlendirilmesindeki ilk yöntem kayıp veri olan kayıtları yok saymaktır. Ancak kayıp verinin çok olduğu yada az kayıtları sahip testlerde bu çözüm sonucu yanlış değerlere saptırmaktadır. Bu yüzden kayıp verilerin yerine bir değer ataması yapmak çok daha iyi bir sonuç olduğu ortaya çıkmıştır. Kayıp verileri Tahmin etmede kullanılan yöntemler aşağıda kısaca anlatılmaktadır.

• Regresyon analizi

Bu yöntem, değerler arasındaki ilişkileri tahmin etmek için kullanılan istatistiksel bir tekniktir. Bu teknik kullanımında önemli nokta, bağımsız değişkenlerin bağımlı değişkenleri açıklama oranının yüksek olması gerekmektedir. Birçok bağımsız değişkenden oluşan bir çoklu regresyon modeli tahmin ettiğimizi varsayalım bu değişkenlerden biri, kayıp gözlemler içeren X olsun. Böyle bir modelde tahmin edilen tüm X gözlemleri diğer bağımsız değişkenler kullanılarak tahmin edilmiştir.

Bu yöntemde tahmin edilen regresyon modeli, kayıp gözlemleri tahmin etmede bir araç olarak kullanılmaktadır. Veriler Tamamen Rassal Olarak kayıp varsayımını sağladığında ve atamalar kayıp gözlem içermeyen diğer bağımsız değişkenlere bağlı olduğunda en küçük kareler katsayıları tutarlıdır yani büyük örneklerde bu işlem neredeyse yansız sonuç verir. En küçük kareler yöntemi, birbirine bağlı olarak değişen iki fiziksel büyüklük arasındaki matematiksel bağlantıyı, mümkün olduğunca gerçeğe uygun bir denklem olarak yazmak için kullanılan, standart bir regresyon yöntemidir.

Bir başka deyişle bu yöntem, ölçüm sonucu elde edilmiş veri noktalarına "mümkün olduğu kadar yakın" geçecek bir fonksiyon eğrisi bulmaya yarar. Şekil 1'de birbiri ile ilişkili iki alana girilen değerlerden oluşacak, regresyon fonksiyonunun oluşturulmasını ve yeni değer tahmin edilmesini sağlar.

Hot Deck

Hot Deck Imputation ile eksik veri değerlerini doldururken benzerlik tahmininde bulunmak için k-en yakın komşu en çok tercih edilen metodudur.

Diğer bir deyişle eksik veri bulduran satır ile tamamlanmış satır arasındaki uzaklık hesabı için k-en yakın komşu metoduyla yapılabilir. Bunun için aşağıdaki adımlar uygulanır.

1.) Veriler tamamlanmış ve tamamlanmamış (eksik) veri kümeleri olmak üzere ikiye bölünür. 2.) X_i tamamlanmış veri kümesinin matrisidir[^] i.durumun j.değişkenini ifade eder.

Y_{ij} tamamlanmamış veri kümesinin matrisidir. Y_{ij} i.durumun j.değişkenini ifade eder.

3.) Her eksik veri içeren satır için öklid uzaklığı hesaplanır.

$$\text{Euclid}(d) = \sqrt{\sum_{j=1}^n (x_{ij} - y_{kj})^2}$$

Uzaklık hesabına göre eksik veri içeren tamamlanmamış satıra en yakın tamamlanmış satır belirlenir.

Hot deck atfının en önemli dezavantajı, 'benzerlik' kavramının tanımlanmasındaki güçlüktür. Bu nedenle hot deck prosedürü kayıp veriler için standart bir yol sağlamamaktadır. Bu benzerliğin belirlenebilmesi için verici (donor) durumların seçimini başarabilecek bir yazılım gerekmektedir. Daha ileri bir hot deck algoritmasına göre, benzer bir kayıttan daha fazla sayıda kayıt belirlenir ve bu verici (donor) kayıtlardan biri kayıp değerlerin atfı için rassal olarak seçilir. Ayrıca eğer uygunsa, bu verici durumların ortalaması kayıp değerlerin atfı için kullanılır.

Son Gözlemi İleri Taşıma

Son Gözlemi İleri Taşıma özellikle uzun süreçli araştırmalarda kullanılan bir yöntemdir. Bu yöntemde

kayıp değer yerine kayıptan önce gözlemlenen son değer atanmasıyla kayıp veri doldurulmuş olur. Bu yöntemde kendisinden önceki veri sonraki eksik verinin yerine konular ve kullanımı oldukça basit ve anlaşılabilir. Bu yöntem değişik zamanlarda alınana sonuçlara göre bir veri kümesi oluşturmada belli bir mantığa sahiptir. Ancak sonuçları açısından da tartışmalı bir yaklaşımdır. •Naive Bayes İle Değer Atama

Naive Bayes sınıflandırıcı Bayes karar teorisine dayanan basit bir olasılıksal sınıflandırıcıdır. Herbir sınıf için olasılıkları hesaplar ve her bir örnek için olasılığı en yüksek sınıfı bulma eğilimindedir. Popüler olmasının sebebi sadece iyi performansı değil basit yapısı yüksek hesaplama hızı ve eksik verilere olan duyarsızlığıdır [5]. NBI Naive Bayesian Classifier kullanan tamamlama metodudur.

Genellikle veritabanlarında kayıp değer taşıyan özellik sayısı 1 den fazla olur. Bu durumda;

1. Tamamlama yapılacak ilk özellik tespit edilmeli.
2. Tamamlanacak özellikler için tamamlanma sıraları göz önünde bulundurulmalı.
- 3 farklı NBI stratejisi vardır.

1.Order Irrelevant Strategy (NBI-OI):

Tamamlanacak özellikler tanımlandıktan sonra veri kümesi eksik değerlerinin tamamlanmasının sırasıyla ilgisi yoktur. Tamamlanmış özelliklerin değerleri daha sonraki özellikler için kullanılmaz. Kayıp değerleri tamamlanmış veri kümeleri tüm farklı tamamlama sıraları için aynıdır.

2. Order Relevant Strategy (NBI-OR):

Tamamlanacak veri kümesi eksik değerlerinin tamamlanma sırasıyla ilgilidir. Tamamlanmış özelliklerin değerleri daha sonraki özellikler için kullanılır. Kayıp değerleri tamamlanmış veri kümeleri tüm farklı tamamlama sıraları için farklıdır.

3. Hybrid Strategy (NBI-Hm):

İlk iki stratejinin birleşimidir. Birinci özellik tamamlama adımında sıralı stratejiyi kullanır kalanında sırasız stratejiyi kullanır.

NBI bu üç stratejiden de anlaşılacağı üzere 2 oluşur:

1.adım tamamlanacak özellikleri ve sırasını tanımlamak. Kayıp değerlere sahip özellikler birden fazla olabilir. Bu kayıp değerleri özellikler arasında önceliğe iki açıdan bakılabilir. Birincisi kayıp değerlerin oranı(missing proportion), ikincisi özelliğin önem faktörü(important factor) dür.

2.adım kayıp veriler için NBI modelini kullanmak. Sıralı stratejide her adımda tamamlanan kayıp değerler ile değişen veri kümesi kullanılır.

Beklenti Maksimizasyonu

EM (Expectation Maximization) Algoritması bir objenin hangi kümeye ait olduğunu belirlemede kesin mesafe ölçütlerini kullanmak yerine tahminsel ölçütleri kullanmayı tercih eder.

Maksimum benzerlik prensibine dayanan Beklenti Maksimizasyonu (BM) algoritması ilk olarak 630

ortaya konulmuştur. Regresyon atamasının iteratif süreçli bir halidir ve 2 iteratif adımdan oluşur.

EM algoritması son yıllarda bir çok araştırmada kullanılan popüler bir yaklaşım olmuştur. EM algoritması, tam olmayan veri problemlerini çözmek için maksimum olasılık tahminlerini yapan tekrarlı bir algoritmadır. EM Algoritmasının her tekrarı iki adımda gerçekleşir. Bu adımlar, bekleneni bulma (E-Adımı) ve maksimizasyon (M Adımı) olarak adlandırılır [2].

E-adımında gözlenen verilerin parametrelerine ait kestirimler kullanılarak bilinmeyen (kayıp) veri ile ilgili en iyi olasılıklar tahmin edilirken, M-Adımında ise tahmin edilen kayıp veri yerine konulup bütün veri üzerinden maksimum olabilirlik hesaplanarak parametrelerin yeni kestirimleri elde edilir [4].

- Karar Ağaçları

C4.5 Karar ağacı algoritması ile kayıp verilerin tahmini yapılabilir. Bunun için aşağıdaki adımlar izlenir;

T çalışılan veri kümesi ve genel bilgi kazancı bilgi(T) olsun. X bu kümenin herhangi bir özelliği olsun ve X özelliğinin bilgi kazancı ise bilgiX(T) olsun. Bilgi(T) hesap edilir. Ancak bilgiX(T) hesap edilirken olmayan veriler bu kümeden çıkarılır. Olay sayısı n ile ifade edilirse ve bilinmeyen veriler b ile ifade edilirse X özelliğinin n-b adet eksik olmayan verileriyle sanki hiçbir veri eksik değilmiş gibi klasik formül uygulanır. Ardından eksik olmayan değerlerin toplam değerlere oranı $F = (n-b) / n$; formülü ile hesaplanır. Bu durumda;

$$F = (\text{bilgi}(T) - \text{bilgiX}(T))$$

formülü ile bilgi kazancı hesaplanmış olur.

Yukarıdaki işlemi tüm eksik veri içeren satırlar için tekrarladığımızda bir tablo elde edilecektir. Elde edilen tabloyu kullanarak her nitelik için geçerli sayıları ve kazanç değerlerini bulabiliriz. Bu tablo ile Karar Ağacı yapısı oluşturulur.

- Çoklu atama

Çoklu atama, Tekli atama yöntemlerinin birleşimini oluşturur. Çoklu atamada Monte Carlo Tekniği kullanılır [2].

$$v = \frac{1}{m} \sum_{i=1}^m \hat{v}_i + \frac{m+1}{m} \left[\frac{1}{m-1} \sum_{i=1}^m (Q_i - \bar{Q})^2 \right]$$

Nokta tahmini için varyans tahmini :

m =Ataması yapılmış ve analiz edilmiş küme sayısı

Q_i = analiz edilmiş i. kümeden tahmin

V_i = analiz edilmiş i. kümeden varyans tahmini

Çoklu atamalardan elde edilen nokta tahmini her analizden elde edilenin

- Yerine ortalamayı koyma

Bu yöntem, veri setinde kayıp verinin olduğu alandaki diğer verilerin ortalamasını alarak kayıp olan verileri doldurmaya yarayan yöntemdir. Veri aralığı aralığı düşük olan verilerde kullanıldığında yararlı olabilir., Aksi halde hata oranını artırır.

3.Sonuç ve Öneriler

Veri madenciliğinde sık ortaya çıkan kayıp veriler, araştırmacılar için dikkate alınmadığında araştırmacıları yanlış sonuçlara götürebilir. Kayıp verilerin tahmininde gözlem sayısı ve verinin özelliği oldukça önemlidir.

Bu çalışmada kayıp veriler ile ilgili sonuçlar aşağıda verilmiştir. Buna göre herhangi bir yöntemin diğerlerinden tamamen üstün olduğu söylenemez. Ancak veri özelliğine, kayıtların birbirleri ile olan ilişkisine, kayıt sayısına ve kayıp veri sayısının toplam kayıt oranına bakarak bir yöntem seçimi yapılabilir.

Her şeyden önce kayıp verileri Durum Düzeyinde Silme işlemi yapılırsa kullanımı basit olmasına rağmen fazla veri kaybında varyans artar ve Rassal Olarak Kayıpta (ROK) hatalı sonuçlar üretir.

Yerine ortalamayı koyma yöntemi, korelasyonun düşmesine ve verilerin dağılımlarını olumsuz olarak değiştirilmesine yol açar. Bu yöntem az kayıp verisi az ve aralığı düşük setlerde kullanılabilir.

Regresyon ataması, bu yöntem için kayıtlarda önce korelasyonu yüksek iki alan seçilip ona göre bir regresyon formülü üretilebilir. İlişkili olmayan alanlarda işe yaramaz. Bu fonksiyonun oluşturulmasında hatanın göz önünde bulundurulması gerekir.

Beklenti maksimizasyonu, Maksimum benzerlik prensibine dayandığı için tüm verilerin kullanılması gerekir. Buradaki benzerlik olabilmesi için ise kayıp verili kayıta benzer yani değer aralığı az olan ve veri seti büyük olan kayıtlarda uygulanması daha doğru sonuçlar verir.

Hot-Deck atama, veriler arasındaki mesafeye bakarak sabit bir sayıyı boş alanlara eklediği için hata bayı oldukça yüksek çıkan bir algoritmadır. Avantajı, uygulamasının kolay olması ve az veri kaybında hatayı fazla etkilememesidir.

Son gözlemi ileri Taşıma, birbirine yakın değerleri olan alanlarda kullanılabilir, ancak veri aralığı yüksek veri setlerinde hatayı yükseltir.

Naive Bayes yöntemi, olasılıksal yöntemleri kullandığı için tüm veriyi kullanır. Bayes bir değerler arasındaki anlamlılığı artırır. Az sayıdaki verilerde hata oranı yüksektir.

Karar ağacı yönteminde kayıp verilerin fazla olması ağaçtaki tutarsızlığı artırmaktadır. Karar ağacında en önemli nokta; onu oluşturan eğitim kümesi ve sağlama kümesi arasındaki ilişkidir. Ağaç karmaşıklaştıkça eğitim kümesi için doğruluğu artmakta, ancak sağlama kümesi için ise doğruluğu azalmaktadır.

Sonuç olarak, elimizde bulunan verinin yapısına ve içeriğine göre algoritmaların farklı problemlerde farklı başarı oranları göstermesi doğaldır. Dolayısıyla en iyi algoritma budur diye genel bir şey yoktur problemin tipine göre kayıp veriyi tespit etme yöntemi değişebilmektedir. Yapay sinir ağları, genetik algoritmalar gibi yapay zeka algoritmaları da kayıp verinin tahmininde kullanılabilirler olup, araştırmalar yapıla bilinir. kümeleme algoritması olduğu için mümkün olduğu kadar fazla veri bulabileceği

4.Kaynaklar

[1] Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M.G., Mallinckrodt, C., Carroll, R.J.: Analyzing Incomplete Longitudinal Clinical Trial Data 2004.

[2] Baygöl, Arzu , Kayıp Veri Analizinde Sıklıkla Kullanılan Etkin Yöntemlerin Değerlendirilmesi,

İstanbul Üniversitesi Sağlık Bilimleri Enstitüsü Yüksek Lisans Tezi 2007.

[3] Dempster A.P., Laird N.M. ve Rubin D.B., . Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 39(1):1-38, 1977.

[4] Bruzzone, L. ve Prieto, F., An Adaptive Semiparametric and Context-Based Approach to unsupervised Change Detection in Multitemporal Remote-Sensing Images. IEEE Transactions on Geoscience and Remote Sensing, 11 (4): 452-466, 2002.

[5] Peng Liu, Lei Lei, Missing Data Treatment Methods and NBI Model, Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06) 2006 IEEE.