YBS Ansiklopedi

www.YBSAnsiklopedi.com

Cilt 4, Sayı 4, Aralık 2017

Öznitelik Mühendisliği (Feature Engineering)

Şadi Evren ŞEKER

1. *İstanbul Şehir Üniversitesi, Yönetim Bilişim Sistemleri Bölümü*

Özet

Bu makalenin amacı, öznitelik mühendisliği kavramını ve bu kavramın güncel kullanımlarını açıklamak, aynı zamanda bu kavramın veri bilimi, iş zekası ve iş analitiği kavramları ile ilişkilerini anlatarak analitik bir sistemin iş süreçlerine etkisi ve sistemin toplam başarısına etkisini ortaya koymaktır. Kavram basitçe, bir makine öğrenmesi veya istatistiksel modelin üzerinde çalışacağı öznitelikleri ortaya çıkarmayı ve bu öz niteliklerin, gerek model gerekse hedeflenen amaç için en iyi şekilde optimize edilmesini amaçlar. Kavram, sahadaki problemlere ve verilere doğrudan dokunduğu için, literatürde genel olarak uygulamalı makine öğrenmesi olarak da geçmektedir.

Anahtar Kelimeler:*feature engineering, machine learning, data mining, data science*

Abstract

The aim of this article is to explain, feature engineering concept and it operational use. Also, the connection between feature engineering and data science, business intelligence and business analytics will be explained in this paper and the affect of feature engineering to the total success of the system. The concept is a methodological approach to extract the features where the machine learning or statistical model will run over and optimize the features for the success of the overall system. The concept is called as applied machine learning in the literature because it directly touches to the field problems and data.

Keywords:*feature engineering, machine learning, data mining, data science*

1. Giriş

Makine öğrenmesi veya istatistiksel modeller, öznitelikler üzerinde çalışır. Problem ve verinin analizinden sonra ve modelin eğitilmesinden hemen önce yer alan öznitelik mühendisliği, iki aşama arasındaki köprüyü oluşturmaktadır ve iki aşamadan da doğrudan etkilenmektedir. Öznitelik mühendisliği, bu modellerin başarısını doğrudan etkileyen özniteliklerin daha başarılı şekilde çıkarılmasını hedefler. Bu hedefe yönelik olarak problem ve veri uzayında, bazı çalışmalar yapar. Bu bölümde, öznitelik mühendisliği çalışmasına örnek olabilecek bazı çalışma örnekleri verilmiştir. Ardından ikinci bölümde bu çalışmaların dayandığı bazı standart yaklaşımlara değinilecek ve literatürdeki bu konudaki güncel yöntemlere değinilecektir. Son olarak öznitelik mühendisliğinin iş analitiği ve yönetim bilişim sistemleri dünyası açısından önemine değinilecektir.

Öznitelik mühendisliği, çok farklı alanlarda kullanılmaktadır. Örneğin, zamana bağlı değerler ve zaman serisi üzerinde öznitelik çıkarılması [1], metin üzerinde veri madenciliği yapılması [2] ve hatta sosyal ağlar üzerinde elde edilen bazı metriklerin öznitelik olarak kullanılması [3] bile literatürde sıkça rastlanan bazı problemleri oluşturur. Hatta bazı çalışmalarda bu özniteliklerin farklı alanlardan çıkarılarak çapraz alanlarda kullanıldığı görülmektedir. Örneğin sosyal ağlar veya internetten elde edilen özniteliklerin şirketlerin borsadaki başarısı ile karşılaştırılması [4]. Bazı veri tipleri üzerinde çıkarılabilecek öznitelikler için standart bazı yöntemler olmasına karşılık, özellikle metin üzerinden öznitelik çıkarılması ve zaman serileri konularında farklı uygulamalar bulunmakta ve bu uygulamaların başarıya etkisi çok çeşitli olabilmektedir. Örneğin zaman serisi üzerine yapılan bir çalışmada, borsa verileri üzerinden farklı teknikler kullanılarak öznitelikler çıkarılmış ve yine bu öznitelikler farklı makine öğrenmesi ve istatistiksel metotları eğitmekte kullanılmıştır. Sonuçta her özniteliğin, her model için farklı etkisi olduğunun anlaşılması yanında zaman serisi için kullanılan yöntemler de bu çalışmada yer almıştır [1]. Giriş kısmında da anlatıldığı üzere, öznitelik mühendisliği, iş analitiği [5], iş analizi veya problem tanımı aşamaları ile, makine öğrenmesi, arasında yer alan öznitelik mühendisliği, bu çalışmada [1] hem ekonomi alanında yayınlanan gazete haberleri üzerinde yapılan metin madenciliği öznitelik çalışmaları hem de bir zaman serisi analizi olan borsa verilerinden öznitelik çıkarma aşamalarında kullanılmıştır.

Metin madenciliği (text mining) [2] olarak geçen kavram, doğrudan metinler üzerinden anlamlı ve sonuca yönelik sonuçlar çıkarılmasını hedefler. Örneğin yazar tanıma problemi, esas olarak yazılan yazıdaki bazı öznitelikleri kullanarak, bu yazıyı kimin yazdığının bulunmasını hedefler. Bu aşamada hangi özniteliklerin kullanılacağı oldukça önemlidir. Örneğin, bir metinden, fiil, isim, sıfat gibi etiketlenen (part of speech tagging) kelimelerin sayısal dağılımı, devrik cümle sayısı, yapılan imla hataları, kullanılan noktalama işaretlerinin sayısı, cümlelerin ortalama uzunluğu gibi çok sayıda sayısal veriye ulaşmak mümkündür ve bu veriler, derlemi oluşturan yazarlar için ayırt edici olabilmektedir. Hatta metinde kullanılan kelimelerin arasında bağlantı kurulması istenen durumlarda, anlambilimsel ağlar (semantic web) veya dizgi mesafe fonksiyonları (string distance functions) kullanılmaktadır [6].

2. Öznitelik Çıkarımının Aşamaları

Öznitelik çıkarımı beş temel adımdan oluşur ve bu adımlar aşağıdaki şekilde sıralanabilir [7]:

- Gösterge Değişkenler (Indicator Variables)
- Etkileşim Öznitelikleri (Interaction Features)
- Öznitelik Gösterimi (Feature Representation)
- Dış Veri (External Data)
- Hata Analizi (Error Analysis)

Yukarıdaki her aşama aşağıda örnekler ile detaylandırılmıştır:

2.1. Gösterge Değişkenler (Indicator Variables)

Öznitelik mühendisliğinin bu aşamasında, amaca yönelik olarak özniteliklerin belirli göstergelere göre şekillenmesi hedeflenir. Örneğin bir sigorta şirketi, araç kaskoları için hedef kitlesine yönelik olarak yaptığı bir pazar araştırmasında, şayet sürücü risklerini hesaplamayı hedefliyorsa, ehliyeti olmayan ve dolayısıyla kasko satışına konu olmayacak olan 18 yaş altı kişileri sistemden eleyebilir. Buradaki 18 yaş sınırı bir eşik değerdir ve gösterge değişken olarak kabul edilebilir. Benzer şekilde satışlar üzerinden yapılacak bir pazarlama analitiği sırasında ramazan ve kurban bayramlarına denk gelen veya bu tarihlerden öncesine denk gelen tarihler için özel gösterge değişkenler tanımlanabilir ve bu gösterge değişkenlere göre özel analizler yapılabilir. Çoklu öznitelikler üzerinden de gösterge değişkenler tanımlamak mümkündür. Örneğin bir müşterinin 18 yaş üzerinde olması ve satış

işleminin bayram öncesine denk gelmesi gibi aynı anda oluşabilecek ve birden fazla değişkene bağlı gösterge değişken tanımlamak mümkündür.

2.2. Etkileşim Öznitelikleri (Interaction Features)

Çıkarılmış olan öznitelikler kullanılarak ve bu öznitelikler üzerinde bazı operatörler kullanılarak yeni özniteliklerin çıkarılması işlemidir. Örneğin iki özniteliğin toplanarak üçüncü ve yeni bir özniteliğin elde edilmesi veya bir öz niteliğin belirli bir değerle işleme sokulması düşünülebilir. Mesela doğum tarihlerinin günümüz tarihinden çıkarılarak yaşın hesaplanması gibi özniteliklerin etkileşime girmesi durumudur. Aşağıda bu konuda kullanılabilecek bazı örnekler verilmiştir:

Toplama: Birden fazla özniteliğin toplanmasıdır. Örneğin satış özniteliklerinin toplanarak toplam satışın bulunması ve bu bilginin yeni bir öznitelik olarak kullanılması düşünülebilir.

Çıkarma: iki öznitelik arasında çıkarma işleminin uygulanmasıdır. Örneğin abonelik başlangıç ve bitiş tarihleri arasındaki farkı alarak abonelik süresinin hesaplanması gibi örnekler düşünülebilir.

Çarpma: iki özniteliğin çarpılması olarak düşünülebilir. Örneğin satış fiyatının hesaplanması için birim fiyat ve satış adedinin çarpılması gibi örnekler düşünülebilir.

Bölme: iki öznitelik arasında bölme işlemi olarak düşünülebilir. Örneğin şubeler arasında müşteri segmentlerinin çıkarılması için, şubenin toplam satışının müşteri sayısına bölünmesi ve şube bazında müşteri başına satış oranının bulunması gibi işlemler düşünülebilir.

2.3. Öznitelik Gösterimi (Feature Representation)

Verinin ifade ettiği anlam, her zaman veri biliminde kullanılan amaca uygun olmayabilir. Aynı verinin farklı amaçlara yönelik olarak farklı gösterimlerinin kullanılması söz konusu olabilir. Örneğin zaman serisi analizi yapılırken, bir satış işleminin tarihinin yanında, haftanın hangi günü olduğu, ayın hangi günü olduğu veya yılın hangi mevsimi olduğu gibi çok farklı özniteliklerin çıkarılması gerekebilir. Genelde tarih ve zamana bağlı bu öznitelik gösterimi dışında, sayısal ve kategorik gösterimler, boş sınıfların gruplanması veya kukla değişken (dummy variable) üretilmesi gibi durumlar da gerekebilir. Örneğin bir kişinin eğitim süresine bakarak ilk, orta veya lise grubuna atanması kategorik bir gösterimdir. Çok düşük frekanslara sahip veya sıfır değerine sahip kayıtların "diğerleri" şeklinde sınıflandırılması boş sınıf gruplaması veya herhangi bir etkiyi işaretlemek için üretilen kukla değişkenler birer öznitelik gösterimi olarak kabul edilebilir.

2.4. Dış Veri (External Data)

Verinin işlenmesi sırasında bazı özniteliklerin, dışarıdan eklenen ilave verilerle farklı özniteliklere dönüştürülmesi mümkündür. Örneğin, bütün eylemler(fiiller) zamana ve konuma bağlı gerçekleşir ve zaman bilgisi olmayan bir aksiyonun zaman bilgisinin kesinleştirildiği ilave özniteliklerin eklenmesi, sistemin başarısına olumlu etki edebilir. Benzer şekilde konum bilgisinin (geocoding) ve olayların gerçekleştiği uzaysal koordinatlar (spatial) arasındaki bağlantılar da önemli olabilir. Bu bilgiler çoğu veri kümesine kolaylıkla eklemlenebilen yapılarda olsalar da çoğu zaman ihmal edilir. Veya sistemde doğrudan kullanılmayan verilerin toplanması da dış veri olarak görülebilir. Örneğin, borsa analizi sırasında ülkenin siyasi hareketliliğinin analiz edilmesi için gazete haberlerindeki olumlu ve olumsuz haber sayılarının sisteme eklenmesi veya hava durumu, döviz kurları gibi bilgilerin sisteme eklenmesi dış verilerin sistem içerisinde kullanılması olarak görülebilir.

2.5. Hata Analizi (Error Analysis)

Model oluşturulduktan sonra, genelde modelin başarısını ölçmek için, oluşturulan ilk model üzerinden hata analizi yapılır. Hata analizi, bu sırada çıkan yüksek orandaki hatanın azaltılması için öznitelik çıkarım aşamasına geri dönülmesi ve özniteliklerin yeniden düzenlenmesi anlamına gelmektedir. Genelde hata analizi sırasında, aşağıdaki yöntemler kullanılabilir:

Büyük hatalarla başlamak. Sistemin başarısını en çok etkileyen durum, genelde hatanın en yüksek olduğu durumdur. Şayet sistemde çok büyük hataya sebep olan durumlar açıklanamıyorsa, mesela anomali veya marjinal durulardan kaynaklandığı açıklanamıyorsa, bu durumda modelin başarısını olumlu yönde etkileyen öznitelik ayarlamalarına gidilebilir.

Hataların segmentlere ayrılması ve veri ile hata arasında açıklanabilir bağlantılar bulunması da diğer bir yöntemdir. Örneğin verinin bazı bölümlerinde yaşanan hataların, doğrudan bölümlerle bağlantısının kurulması ve sonrasında bu bölümler için özel öznitelik mühendisliği yapılması (mesela kukla değişken kullanılması, veya öznitelik etkileşime girilmesi birer yöntem olabilir). Örneğin, kredi skorunu hesaplayan bir makine öğrenmesi modeli düşünelim, bu modeldeki hatanın düşürülmesi için, yaş ile modeldeki hata arasında ilişki bulmak ve mesela 30 yaş altı, 30 – 50 yaş arası veya 50 yaş üzeri segmentlerindeki hataların farklı davrandığının bulunması ve hatta bu şekilde 3 farklı segment oluşturarak her segment için farklı makine öğrenmesi algoritmasının önerilmesi bir yöntem olabilir.

Genelde hataların segmentasyonu ile ilgili sıkıntı yaşandığı durumlarda, mesela segmentler ve hata arasındaki bağlantının kurulamadığı durumlarda, gözetimsiz öğrenme (unsupervised learning) yöntemlerini kullanarak bu segmentlerin belirlenmesi mümkündür. Yine bu bağlantıların kurulması sırasında, alan bilgisi olan uzman görüşleri de faydalı olabilmektedir.

3. Öznitelik Çıkarım Süreci ve Tehditler

Çoğu veri madenciliği projesinde, benzer sürecin tekrar etmesi sonucunda, öznitelik çıkarım süreci bir liste halinde sıralanabilir. Genelde bu süreç aşağıdaki şekilde ilerler:

1. Öznitelikler üzerinde beyin fırtınası yapıldığı, çıkan özniteliklerin test edildiği, daha önce çıkan özniteliklerle aralarında korelasyon testlerinin yapıldığı ve elde bulunan özniteliklerin tanındığı süreçtir.
2. Hangi özniteliklerin türetileceğine karar verilen ikinci aşamada, problem ve olası modeller değerlendirilerek sistemin başarısını olumlu etkileme potansiyeli olan öznitelikler ortaya konur.
3. Özniteliklerin üretilme aşamasıdır.
4. Bir önceki adımda karar üretilen özniteliklerin model ile ne kadar uyumlu çalıştığına bakılır.
5. Gerekli olması durumunda birinci veya ikinci adıma geri dönülerek öznitelikler yeniden gözden geçirilir.

Öznitelik çıkarım sürecinde yaşanan diğer bir konu ise, öznitelik seçimi aşamasıdır. Yani, öznitelik mühendisliği, genelde özniteliklerin türetilmesini ve gerek eldeki mevcut öznitelikler üzerinde çalıştırılan işlemler, gerekse dış verilerle desteklenen işlemler sonucunda özniteliklerin mühendisliğinin yapıldığı ve üretildiği veya iyileştirildiği süreçtir. Bu işlem, öznitelik seçimi (feature selection) ile karıştırılmamalıdır. Genelde öznitelik seçimi, öznitelik mühendisliğinden farklı işler ve öznitelik seçiminde, herhangi yeni bir öznitelik üretilmez veya mevcut öznitelikler üzerinde herhangi bir değişiklik yapılmaz.

Öznitelik mühendisliğinde, dikkat edilmesi gereken diğer bir durum ise, öznitelik patlamasıdır. Öznitelik üretim süreci otomatik bir yapıya bağlandığında veya çok fazla öznitelik üretimi denendiğinde, işlem kapasitesinin üzerinde

ve süreci yavaşlatan ve hatta bazı durumlarda sistemi hataya götüren öznitelik sayılarına ulaşılma riski bulunur. Bunun için öznitelik patlamasını durdurucu önlemler alınabilir. Örneğin öznitelik seçimi (feature selection), bu aşamada devreye girebilir.

Yine öznitelik sayısının artması durumuna karşı, özniteliklerin boyut indirgeme (dimension reduction), boyut dönüşümü, öznitelik karımlama (feature hashing) veya aşırı öğrenme (overfitting) için kullanılan düzenlileştirme (regularization) gibi yöntemler de kullanılabilir.

4. Öznitelik Mühendisliği Örnekleri

Bu bölümde, farklı veri ve problem kümelerindeki öznitelik çıkarım örneklerine yer verilecektir.

4.1. Zamansal Öznitelik Kazanımı

Özellikle zamana bağlı tahminler, zaman serisi analizi veya aksiyon/eylem/fiil temelli özniteliklerin tamamında zaman, unutulmaması gereken önemli bir özniteliği oluşturur. Zamansal öznitelik kazanımında, zamana göre özniteliğin etkisini formüllendirmek hedeflenir [8]. Bir örnek üzerinden zamansal öznitelik kazanımını açıklamaya çalışalım. Örneğin bir kafede satılan kahvenin satış tahminlerini yapmak istediğimiz bir modelde. Kafedeki geçmiş satışları analiz eden bir model oluşturuluyor ve bu modelin üzerindeki zaman etkisi inceleniyor olsun. Böyle bir durumda, aşağıdaki farklı durumların kombinasyonlarının dikkate alınmasında fayda vardır.

1. Son 12 saat içerisindeki kahve servislerinin sayısı
2. 24 saat önce (aynı saat dilimindeki) kahve servislerinin sayısı
3. 7 gün önce (haftanın aynı günü) kahve servislerinin sayısı
4. 365 gün önce (geçen sene aynı gün) kahve servislerinin sayısı
5. Tatil ve haftanın günü etkisi (hafta sonu / hafta içi)
6. Hava durumu (bu bir dış veri olarak kabul edilebilir)

Yukarıdaki özniteliklerin dışında bu özniteliklerin birlikteliğinden de yeni öznitelikler çıkarılabilir. Örneğin 2,3 ve 4 numaralı öznitelikler tekrar eden (recurring) [9] ve sezonsal etkilerin bulunması için kullanışlı olabilir.

4.2. Metin Üzerinde Öznitelik Mühendisliği

Metin üzerinde anlam çıkarımı, aslında çoğu makine öğrenme ve istatistiksel model için anlamsız olan metinlerin işlenerek anlamlı sayılara dönüştürülmesi sürecidir. Örneğin, metinde geçen kelime sayıları, çoğu problem için sıkça başvurulan bir değerdir. Genelde bu kelime sayılarından oluşturulan bir diziye vektör ismi verilir ve kelime frekansına bağlı olarak vektörler üzerinde farklı yöntemlerle öznitelik çıkarımı kullanılabilir. Örneğin terimlerin frekansı ve aynı zamanda terimlerin geçtiği dokümanların frekanslarını hesaplayan TF-IDF gibi değerlerin çıkarılması mümkün olur. Benzer şekilde, çoğu derin öğrenme algoritmaları için kelimelerin vektöre dönüşümü (word2vec) oldukça kritik sonuçlar doğurabilmektedir.

4.3. Resim Öznitelikleri

Resimler üzerinde öznitelik çalışmaları, farklı sinyal işleme algoritmaları veya görüntü işleme yöntemlerinin birlikte kullanılması ile ortaya çıkmaktadır. İki boyutlu bir resmi aslında iki boyutlu bir dizi olarak düşünmek mümkündür ve bu dizi üzerinde başta kayan pencereler olmak üzere çok farklı yöntemlerle öznitelikler çıkarılabilir.

Örneğin, resim işleme sırasında, sıkça başvurulan histogram of gradients (HOG) yöntemi, resmi yatay ve dikey olarak parlaklık akışlarına bölmekte ve bu noktalarda, resmi oluşturan kenar, kontur veya dokular bulunabilmektedir. Bu yöntem, sonuçta her hücre için tek boyutlu bir öznitelik dizisi çıkarmakta ve bu öznitelik dizisi kullanılarak yüz tanıma gibi işlemler yapılabilmektedir.

5. Öznitelik Mühendisliğinin Önemi

Yukarıda anlatılan öznitelik çalışmaları, veri bilimi için çok kritik ve işlenebilir verilerin elde edilmesi aşamasında kullanılır. Çoğu durumda, eksik verilerin [10] çözümünden normalde işlenmesi güç olan metin işleme, resim işleme gibi problemlere kadar çok farklı problemlerin çözülmesine imkan vermesi, verinin zamana bağlı bir seri olarak kullanılabilmesi gibi çok farklı avantajları vardır. Çoğu durumda, öznitelik çıkarımının bir akış şeklinde (pipeline) birbirine bağlı aşamalardan oluştuğunu söylemek de mümkündür.

Öznitelik mühendisliğini, genel olarak sistemin başarısını arttıran bir süreç olarak görmek mümkündür. Ancak bu yaklaşımın, yanlış yönlendirme ihtimali bulunur. Genelde, ulaşılan sonuçlar, seçilen model ve özniteliklerin bir neticesidir ve iyi sonuçlar, her zaman için iyi bir veri madenciliği sürecine işaret etmez. Örneğin iyi bir öznitelik mühendisliği sonucunda ulaşılan doğru öznitelikler, daha basit modellerin daha başarılı çalışmasına imkan sağlar. Basit modeller ise daha hızlı çalışan, basit anlaşılan ve basit şekilde sürdürülebilen sistemlerin inşa edilmesi için çok önemlidir. Bu açıdan öznitelik mühendisliğinin, sistemin esnekliğine katkı sağladığı söylenebilir.

Kaynakça

[1] Seker, Sadi Evren, et al. "Ensemble classification over stock market time series and economy news." Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on. IEEE, 2013.

[2] Seker, Sadi Evren. "Metin Madenciliği (Text Mining)." YBS Ansiklopedi 2.3 (2015): 30-32.

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

[4] Seker, Sadi Evren, Bilal Cankir, and Mehmet Emin Okur. "Strategic Competition of Internet Interfaces for XU30 Quoted Companies." International Journal of Computer and Communication Engineering 3.6 (2014): 464.

[5] Seker, Sadi Evren. "İş Analitiği (Business Analytics)." YBS Ansiklopedi (2016).

[6] Seker, Sadi Evren, et al. "A novel string distance function based on most frequent K characters." arXiv preprint arXiv:1401.6596(2014).

[7] Elite DataScience, Best Practices for Feature Engineering, (2016)

[8] Seker, Sadi Evren. "Temporal logic extension for self-referring, nonexistence, multiple recurrence, and anterior past events." Turkish Journal of Electrical Engineering & Computer Sciences 23.1 (2015): 212-230.

[9] Al-Khateeb, Tahseen, et al. "Recurring and novel class detection using class-based ensemble for evolving data stream." IEEE Transactions on Knowledge and Data Engineering 28.10 (2016): 2752-2764.