İstatistikçiler Dergisi: İstatistik & Aktüerya

Journal of Statisticians: Statistics and Actuarial Sciences

IDIA 10, 2017, 1, 1-10

Geliş/Received:01.09.2016, Kabul/Accepted: 23.06.2017

www.istatistikciler.org

Araştırma Makalesi / Research Article

Bayesci ağ yapısının öğrenilmesinde grafiksel bir yaklaşım

Yasemin Kayhan Atılgan

Hacettepe Üniversitesi
İstatistik Bölümü
06800-Beytepe, Ankara, Türkiye
ykayhan@hacettepe.edu.tr

Derya Ersel

Hacettepe Üniversitesi
İstatistik Bölümü
06800-Beytepe, Ankara, Türkiye
dtektas@hacettepe.edu.tr

Öz

Bu çalışmada, uzman bilgisi olmadığında Bayesci ağ yapısının öğrenilmesinde, çok boyutlu veri kümesindeki değişkenler arasındaki ilişki yapısını ortaya koyan Robust Coplot [1] grafiğinden yararlanılması önerilmiştir. Böylece, zaman alıcı öğrenme algoritmalarına gerek kalmadan Bayesci ağın oluşturulması sağlanmıştır. Önerilen yöntem bir veri kümesi üzerinde uygulanarak sonuçlar tartışılmış ve Robust Coplot grafiği ile veri kümesinin ön incelemesinin uzman bilgisi eksikliğini büyük ölçüde giderdiği gösterilmiştir.

Anahtar sözcükler: Bayesci ağlar, Bayesci ağlarda öğrenme, Robust Coplot.

Abstract

A Graphical approach to learning Bayesian networks

In this study, it is proposed to use Robust Coplot [1] that reveals the relationship structure between variables in multi-dimensional dataset to learn Bayesian network structure in the absence of expert knowledge. Hence, it is provided to create Bayesian networks without the need for time-consuming learning algorithms. The proposed method is applied to a data set and the results are discussed. Besides, it is shown that preliminary examination of data set with Robust Coplot provides to fulfill the deficiency of expert knowledge to a large extent

Keywords: *Bayesian networks, learning Bayesian Networks, Robust Coplot.*

1. Giriş

1990'lı yıllarda kullanılmaya başlanan Bayesci ağlar, çok boyutlu veri kümesindeki rastlantı değişkenleri arasındaki olasılıksal ilişkileri kodlayan grafiksel modellerdir. Hem nedensel hem de olasılıksal özelliklere sahip olduklarından, bu ağlar ile veri bilgisi ve uzman görüşü kolaylıkla birleştirilebilir. Bayesci ağlar ile ayrıca, ilgilenilen problemin kesin olmayan tanım bölgesi ile ilgili bilgi temsil

edilebildiği gibi, güçlü çıkarsamalar da yapılabilir. İstatistiksel analizlerde Bayesci ağlardan yararlanmak kullanıcıya birçok üstünlük sağlar. Bu üstünlüklerden bazıları; değişkenler arasındaki nedensel ilişkilerin anlaşılmasını sağlamaları, olasılık kuramına dayandığından her zaman tutarlı sonuçlar vermeleri, robust olmaları, uzman görüşünü modellemeye katmaları ve veride kayıp gözlem olması durumunda da güvenilir çıkarsamalar yapmaları olarak verilebilir [8, 12].

Bayesci ağlar, istatistik, makine öğrenmesi ve yapay zeka alanlarında çok kullanılan ve "yönlü dönüşsüz grafik (directed acyclic graph [DAG])" olarak bilinen bir grafiksel model yapısına sahiptir. Sezgisel olarak anlaşılabilir bir yapıya sahip olan bu ağlar, bir rastlantı değişkenleri kümesinin çok değişkenli olasılık dağılımının etkili bir gösteriminin ve bu gösterim üzerinden çeşitli hesaplamaların yapılmasını sağlar [2]. Son yıllarda Bayesci ağlar, uzman sistemlerinde, kesin olmayan uzman görüşlerinin sisteme girmesini sağlayan önemli bir yöntem olarak karşımıza çıkmaktadır.

Diğer olasılıksal grafik modellerinin aksine, Bayesci ağlardaki tüm parametrelerin anlaşılabilir bir yorumu vardır. Bu nedenle Bayesci ağlar, zaman kaybettiren öğrenme süreçlerine gerek kalmadan uzman görüşü kullanılarak doğrudan oluşturulabilir. Uygun uzman görüşüne ulaşmak her zaman mümkün olmadığından Bayesci ağların oluşturulmasında veri kümesinden yararlanmak gerekmektedir [8, 12]. Bu konu "Bayesci ağlarda öğrenme problemi" olarak bilinir ve veri ile önsel bilgi (uzman görüşü, nedensel ilişkiler) verildiğinde ağ yapısının ve parametrelerin tahmin edilmesi olarak tanımlanır. Öğrenme, çok canlı bir araştırma alanıdır ve bu alanda birçok yöntem geliştirilmektedir. Ağ yapısının öğrenilmesi, parametrelerin öğrenilmesinden daha zor bir problemdir. Ayrıca, gizli düğümler ya da kayıp veri gibi kısmi gözlenebilirlik durumu söz konusu olduğunda başka zorluklar da ortaya çıkar [2,13].

Bu çalışmada, Bayesci ağ yapısının öğrenilmesi problemi üzerinde durulmuştur. Literatürde, Bayesci ağ yapısının öğrenilmesi ile ilgili çalışmalar Bayesci yöntemlere dayandığı gibi, yarı Bayesci ya da Bayesci olmayan yöntemlere de dayanabilir. Bu yöntemlerin çoğu, skor ölçüsü ve arama işlemi olmak üzere iki bileşene dayanır. Skor ölçüsü, veri ile bir ağ yapısı arasındaki uyum iyiliğini ölçer. Arama işlemi ile bu skor ölçüsünden yararlanılarak değerlendirilecek ağ yapıları oluşturulur [5]. Skor ölçülerine dayalı bu yaklaşımların dışında Bayesci ağ yapısının belirlenmesinde ya da güncellenmesinde diğer istatistiksel ve veri madenciliği yöntemlerinin sonuçlarından da yararlanılabilir. Özellikle uzman bilgisinin yokluğunda ya da yeterli olmadığı durumlarda diğer yöntemlerden elde edilen sonuçlar önsel bilgi olarak kullanılabilir. Örneğin, Dong-peng ve Jin-lin (2008) [7] çalışmalarında bireylerin banka kredi değerlendirmesinde kullanılacak bir Bayesci ağ modelinin belirlenmesinde birliktelik analizinden elde edilen sonuçlardan yararlanmışlardır. Bu çalışmada ise önsel bilginin olmadığı durumda Bayesci ağ modeli belirleme de çok boyutlu veri kümesinin iki boyutlu düzlemde bir grafiğinin sunan Robust Coplot yönteminden yararlanılmıştır.

Çok boyutlu veri kümesinin grafiksel olarak gösterimi, 1940 ve 1960 yılları arasında çok yoğun çalışılmış ve hala istatistikçiler tarafından yoğun olarak çalışılan konulardan biridir [20]. Özellikle mühendislik, tıp, ekonomi ve istatistik gibi alanlardaki çalışmalarda veri kümesine ileri istatistiksel analizler uygulanmadan önce, grafiksel olarak ön incelemesini yapmanın, veri kümesinin yapısını anlamanın birçok avantajı vardır [20]. Literatürde bu amaçla kullanılan temel bileşenler analizi, kümeleme analizi, diskriminant analizi gibi birçok çok değişkenli istatistiksel yöntem yer almaktadır. Coplot yöntemi çok boyutlu veri kümesinde değişkenler arasındaki, değişkenler ve gözlemler arasındaki ve gözlemlerin kendi aralarındaki ilişki yapılarını iki boyutlu tek bir grafik üzerinden inceleyebilmek amacıyla literatürde farklı alanlarda, çeşitli amaçlar için kullanılmıştır. Coplot yönteminin bahsedilen yöntemlerden farkı, aralarında yüksek korelasyona sahip değişkenlerin bir araya gelerek oluşturduğu yeni bir bileşik değişken üzerinden yorumlama yapmak yerine, doğrudan veri kümesinin orijinal değişkenlerini resmetmesidir [4]. Ayrıca bu yöntem, değişkenler arasındaki ya da gözlemler arasındaki ilişki yapısını ayrı ayrı değil tek bir grafik ile sunar. Yöntemin dezavantajı, veri kümesi aykırı değer içerdiğinde grafiksel gösterim üzerinde bozulmalar yaşanmasıdır. Bu sorunu ortadan kaldırmak ve çok boyutlu veri kümesi aykırı değer içerdiğinde bile güvenilir sonuçlar üretebilmek amacıyla Robust Coplot yöntemi geliştirilmiştir [1].

Literatürde Coplot yöntemi, şehirler arasındaki sosyoekonomik farkların incelendiği [14], bilgisayar bilimlerinde paralel iş yükü modellerinin karşılaştırıldığı [19], coğrafi bir bölgenin iklim kuşaklarının

belirlendiği [10], sağlık hizmetlerinde çok boyutlu veri kümesinin incelendiği [4] çalışmalarda grafiksel gösterim aracı olarak kullanılmıştır. Yine bu yöntemden veri zarflama analizi sonuçlarının grafiksel gösterimini gerçekleştirmek amacıyla da faydalanılmaktadır [17]. Ayrıca çok boyutlu veri kümesindeki olası aykırı değerleri ya da etkili gözlemleri tespit etme aracı olarak da kullanılabilmektedir [15]. Benzer şekilde aykırı değerlerden etkilenmeyen Robust Coplot yöntemi güneş enerjisi modelleme çalışmalarında çoklu bağlantı sorununu ortadan kaldırmak amacıyla değişken seçim aracı olarak da kullanılmıştır [6]. Sonuç olarak, çok boyutlu veri kümesinin grafiksel ön incelemesinin faydalı araştırmacıya avantajlar sunacağı birçok istatistiksel analiz yöntemi vardır.

Bu çalışmada çok boyutlu veri kümesinin Robust Coplot yöntemi kullanılarak ön incelemesi yapılmış ve elde edilen grafik sayesinde değişkenler arasındaki ilişkiler ortaya çıkartılmıştır. Uzman görüşü olmadığında, Bayesci ağlardaki değişkenler arasındaki bağlar, Robust Coplot grafiğinden elde edilen ilişkilere göre oluşturulmuştur. Böylece, önsel bilgi yokluğunda veriye uygun bir ağ yapısının elde edilmesi amaçlanmıştır. Çalışmanın ikinci bölümünde, Robust Coplot yöntemi ve Bayesci ağlar tanıtılmış, Bayesci ağ yapısının oluşturulmasında Robust Coplot yönteminin nasıl kullanıldığı açıklanmıştır. Üçüncü bölümde, bu iki yöntemin bir arada kullanılmasının Bayesci ağ yapıları oluşturmada araştırmacıya sağlayacağı avantajlar bir örnek üzerinden sayısal olarak sunulmuştur. Sonuç bölümünde, önsel bilginin olmadığı durumlarda Bayesci ağ yapısı oluşturmadan önce çok boyutlu veri kümesinin grafiksel ön incelemesinin yapılmasının, araştırmacıya sağlayacağı kolaylıklar üzerinde durulmuş ve ileride yapılacak çalışmalar için önerilerde bulunulmuştur.

2. Yöntem

Bu bölümde Robust Coplot yöntemi ile Bayesci ağlar kısaca tanıtılacak ve bu iki yöntemin Bayesci ağ yapısının öğrenilmesinde bir arada nasıl kullanılacağı açıklanacaktır.

2.1. Robust Coplot

Robust Coplot yöntemi temelde iki aşamadan oluşmaktadır [1]. İlk aşamada nxp boyutlu veri kümesindeki gözlemlerin iki boyutlu düzlemdeki gösterimi robust çok boyutlu ölçekleme yöntemi [9] kullanılarak elde edilir. İkinci aşamada ise gözlemler ile veri kümesindeki değişkenler arasındaki medyan mutlak sapma korelasyon katsayısı [18] ile hesaplanan ilişki miktarlarından yararlanılarak grafiğe değişkenleri temsil eden vektörler uygun biçimde yerleştirilir. Bu iki aşama üç adımda aşağıdaki gibi gerçekleştirilir [1].

Birinci adımda, X_{nxp} boyutlu veri matrisi Eşt. 1 ile Z_{nxp} standartlaştırılmış veri matrisine dönüştürülür.

$$z_{ij} = \frac{x_{ij} - med\left(\mathrm{x}_j\right)}{MAD\left(\mathrm{x}_j\right)}, \quad i = 1,2,...,n \quad j = 1,2,...,p. \tag{1}$$

Burada, z_{ij} standartlaştırılmış veri matrisinin i. satır, j. sütun elemanı x_j veri matrisinin j. sütunu, $med(.)$ medyan fonksiyonu, $\mathrm{MAD}(.)$ ise medyan mutlak sapma fonksiyonu olarak tanımlanmaktadır.

İkinci adımda, standartlaştırılmış veri kümesi kullanılarak gözlemlerin iki boyutlu düzlemdeki gösterimi için Eşt. 2 kullanılır.

$$f\left(O,Y\right) = \sum_{i<j}\left[\delta_{ij} - d_{ij}\left(Y\right) - o_{ij}\right]^2 + \lambda\sum_{i<j}\left|o_{ij}\right|. \tag{2}$$

Burada, δ_{ij} standartlaştırılmış veri matrisinin i. ve j. satırları arasındaki benzemezlik metriği, Y_{nx2} iki boyutlu düzlemdeki koordinat matrisi, $\lambda > 0$ veri kümesindeki aykırı değer miktarını belirleyen kontrol parametresi, o_{ij} de, aykırı değer matrisinin i. satır, j. sütun elemanı olarak tanımlanmaktadır.

Son adımda, ikinci aşamada elde edilen grafiğin üzerine her bir değişkene karşılık gelen bir vektör eklenir. Yerleştirilecek vektörlerin doğrultusuna ve büyüklüğüne Eşt. 3 ve Eşt. 4 kullanılarak karar verilir.

$$r_j = \frac{MAD^2\left(u_j\right) - MAD^2\left(k_j\right)}{MAD^2\left(u_j\right) + MAD^2\left(k_j\right)}, \tag{3}$$

burada, u_j ve k_j Eşt. 4 ile verilen robust temel değişkenlerdir.

$$\begin{aligned} u_j &= \frac{z_j - med\left(z_j\right)}{MAD\left(z_j\right)} + \frac{v_j - med\left(v_j\right)}{MAD\left(v_j\right)}, \\ k_j &= \frac{z_j - med\left(z_j\right)}{MAD\left(z_j\right)} - \frac{v_j - med\left(v_j\right)}{MAD\left(v_j\right)}. \end{aligned} \tag{4}$$

Burada, z_j standartlaştırılmış veri matrisinin j. sütunu, v_j ise robust çok boyutlu ölçekleme grafiğindeki n tane gözlemin belirli bir doğrultuda j. değişkeni temsil eden vektör üzerine iz düşüm değerleridir.

Bu üç adımın sonunda elde edilen Robust Coplot grafiğinde, birbiri ile benzerlik gösteren gözlemler birbirlerine daha yakın, farklılık gösteren gözlemler ise daha uzak yerleşecektir. Bu sayede gözlemler arasındaki ilişki görsel olarak incelenebilir. Ayrıca veri kümesinin çoğunluğundan uzakta yerleşecek bir gözlem şüpheli gözlem olarak düşünülür ve olası aykırı değer olarak incelenebilir. Birbirleri ile aynı yönde ve yakın yerleşen vektörler, karşılık gelen değişkenlerin arasındaki ilişkinin yüksek olduğunu ifade eder. Birbiri arasındaki açı 90 dereceye yakın olan iki vektör, karşılık gelen değişkenlerin ilişkisiz olduğunu ifade eder. Aynı zamanda belli bir grup gözlem doğrultusunda yerleşen bir vektör, ilgili gözlemlerin o değişkene ilişkin değerlerinin yüksek olacağını ifade eder [1].

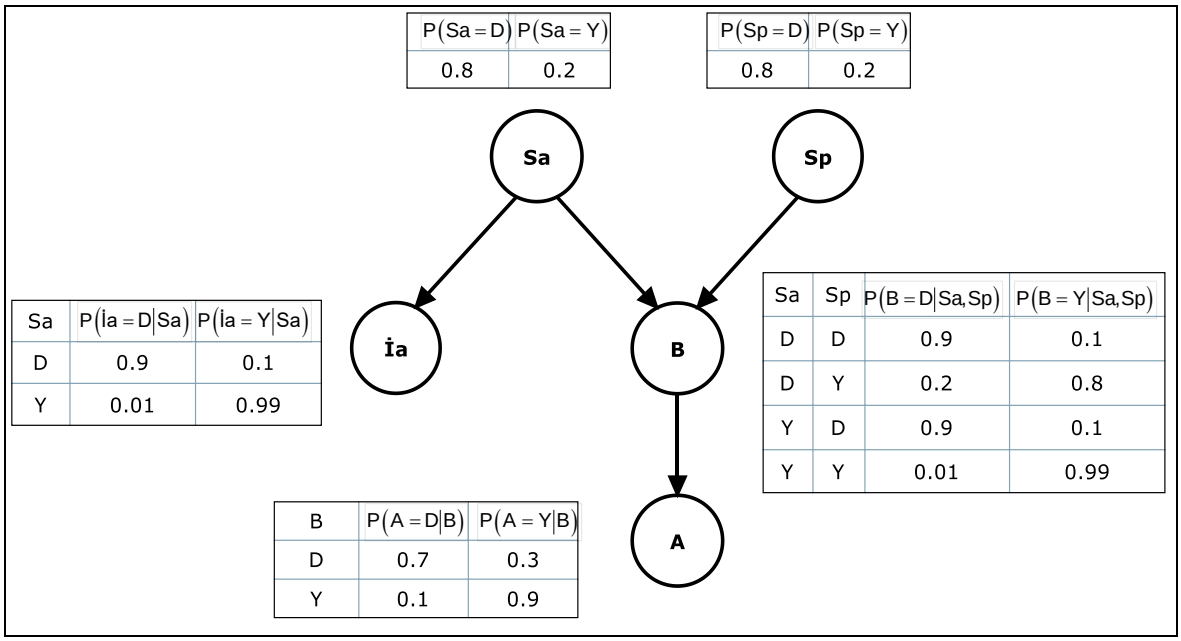
2.2. Bayesci ağlar

Bir Bayesci ağ, X_{nxp} veri matrisindeki p tane rastlantı değişkenine ilişkin çok değişkenli olasılık dağılımını temsil eden bir DAG'dır. Ağ, G ve Θ olmak üzere iki bileşenden oluşur ve $BN = \left(G, \Theta\right)$ biçiminde gösterilir. İlk bileşen G, düğümlerin $X_1, \ldots, X_p$ rastlantı değişkenlerini, düğümler arasındaki bağların ise bu değişkenler arasındaki doğrudan bağımlılıkları gösterdiği bir grafik yapısıdır. G grafiği koşullu bağımsızlık varsayımlarını içerir. Bayesci ağların ikinci bileşeni Θ, ağdaki parametrelerin kümesini gösterir. Bu parametreler, Bayesci ağdaki her bir X_j rastlantı değişkenine ilişkin koşullu olasılık dağılımlarıdır. Bir X_j rastlantı değişkeni için koşullu olasılık dağılımı, X_j'nin G'deki ebeveynlerinin kümesi π_j olmak üzere, $\theta_{x_j|\pi_j} = P_{BN}\left(X_j \middle| \pi_j\right)$ biçiminde tanımlanır. Bir rastlantı değişkeninin verilen Bayesci ağda bir ebeveyni yoksa bu rastlantı değişkeni için koşullu olasılık dağılımı, marjinal olasılık dağılımına karşılık gelir. Bu parametrelerden ve Bayesci ağ yapısından yararlanarak, p tane rastlantı değişkeni için tek bir çok değişkenli olasılık dağılımı tanımlanır ve bu dağılım Eşt. 5'ten

yararlanılarak elde edilir. Çok değişkenli olasılık dağılımının bu eşitlikten elde edilmesi "zincir kuralı (chain rule)" olarak adlandırılır [3, 13].

$$P\left(X_1,\ldots,X_n\right)=\prod_{i=1}^{n}P\left(X_i\middle|\pi_i\right)=\prod_{i=1}^{n}\theta_{X_i|\pi_i} \tag{5}$$

Bayesci ağların ve bazı özelliklerinin daha iyi anlaşılması için bir örnek üzerinde durulsun [2]. Bir kişinin belinin incinmesine neden olan olaylar incelenmek istensin. Bel incinmesi olayı "Bel (B)" değişkeni ile gösterilsin. Bel incinmesi bel ağrısına neden olabilir. Bel ağrısı, "Ağrı (A)" değişkeni ile gösterilsin. Bel incinmesinin nedeni yanlış yapılan spor aktivitesi olabilir. Bu olay, "Spor (Sp)" değişkeni ile temsil edilsin. Diğer bir neden ise, kişinin iş yerinde oturduğu sandalyenin konforsuz olması olabilir. Kişinin sandalyesinin konforsuz olması durumu, "Sandalye (Sa)" değişkeni ile gösterilsin. Bu neden söz konusu olduğunda, bu kişinin iş arkadaşlarının da benzer bir bel problemine sahip olup olmadıkları da araştırılabilir. Bu durumda ilgili değişken "İş Arkadaşı (İa)" olarak alınabilir. Bu problemdeki tüm değişkenler iki düzeyli ve bu düzeyler "Doğru (D)" ve "Yanlış (Y)" biçimindedir. Bu problem için oluşturulan Bayesci ağ Şekil 1 ile verilmiştir.



Şekil 1. Sırt incinmesi örneği için Bayesci ağ.

Şekil 1'de tüm düğümler için koşullu olasılık tablosu (conditional probability table – CPT), ilgili düğümün yanında verilmiştir. Bayesci ağların koşullu bağımsızlık özelliği, zincir kuralından yararlanarak değişkenlerin çok değişkenli olasılık dağılımının daha basit bir şekilde ifade edilmesini sağlar. Örneğin, Şekil 1'e göre

$$P\left(Sa,Sp,İa,B,A\right)=P\left(Sa\right)P\left(Sp\right)P\left(İa\middle|Sp,Sa\right)P\left(B\middle|İa,Sp,Sa\right)P\left(A\middle|B,İa,Sp,Sa\right)$$

biçiminde yazılan çok değişkenli olasılık dağılımı, koşullu bağımsızlık özelliğinden yararlanılarak

$$P\left(Sa,Sp,İa,B,A\right)=P\left(Sa\right)P\left(Sp\right)P\left(İa\middle|Sa\right)P\left(B\middle|Sp,Sa\right)P\left(A\middle|B\right)$$

olarak yazılabilir. Böylece, modeldeki parametre sayısı $2^5 - 1 = 31$'den 10'a düşer. Parametre sayısında bu azalış modeldeki çıkarsamaların, hesaplamaların ve öğrenmenin gerçekleştirilmesinde büyük kolaylıklar sağlar. Daha az parametreye sahip olan bu model yanlılık ve varyans etkilerine karşı daha robusttır [2].

Bayesci ağlarda öğrenme, parametrelerin öğrenilmesi ve ağ yapısının öğrenilmesi olarak ikiye ayrılır. Bu çalışmada Bayesci ağ yapısının öğrenilmesi üzerinde durulmuştur. Veri kümesi hakkında uzman bilgisi olmadığında diğer istatistiksel yöntemler yardımıyla değişkenler arasındaki ilişki yapısının bir ön incelemesi yapılarak Bayesci ağdaki değişkenler arasındaki bağlar oluşturulabilir. Bu çalışmada, Bayesci ağ yapısının öğrenilmesi için Bölüm 2.1'de tanıtılan, çok boyutlu verinin görsel olarak incelenmesinde kullanılan Robust Coplot yönteminden yararlanılmıştır. İlk olarak, ilgilenilen çok boyutlu veri kümesi ile ilgili Bayesci ağ uzman bilgisi olmadan, yalnızca araştırmacının önsel bilgisine bağlı olarak oluşturulmuştur. Daha sonra, aynı veri kümesine Robust Coplot yöntemi uygulanmış ve değişkenler arasındaki ilişki yapısı tek bir grafik ile özetlenmiştir. İlk oluşturulan Bayesci ağdaki bağlar, elde edilen Robust Coplot grafiğinden yararlanılarak güncellenmiştir. Güncelleme, ilk Bayesci ağa yeni bağların eklenmesi, ilk ağdan mevcut bağların çıkartılması ya da bağ ekleme ve bağ çıkarma işlemlerinin birlikte yapılması şeklinde gerçekleştirilebilir. Bu güncellemeler sonucu oluşturulan yeni Bayesci ağın eski bağa göre veri kümesini daha iyi temsil edip etmediğine ise bazı ölçütler ile değerlendirilmiştir.

Bayesci ağlarda, gerçek gözlemlerin bir kümesi üzerinden tahminlerin ve tanıların gerçek durumlara uyumunu ölçerek Bayesci ağı derecelendiren ölçütler vardır. Bu ölçütler ile ağdaki zayıf noktalar belirlenebildiği gibi gerçek durumlarla zayıf ilişkili tahmin yapan düğümler de belirlenebilir. Böylece, bu düğümlerin koşullu olasılık tablolarının yeniden oluşturulması, öğrenme verisinin artırılması ya da ağın güncellenmesi gibi değişiklerle tahminlerin performansları artırılabilir.

Ölçütler elde edilirken ağdaki gözlemler "gözlenmiş" ve "gözlenmemiş" olarak ikiye ayrılır. Gözlenmiş düğümlerin değerleri veri dosyasından çekilir. Gözlenmiş değerler daha sonra Bayesci inanç güncellemesi ile gözlenmemiş düğümlerin değerlerini tahmin etmede kullanılır. Bu süreç, veri dosyasındaki her bir durum için tekrar edilir. Her bir durum için gözlenmemiş düğümlerin tahmin değerleri veri dosyasındaki gerçek gözlenen değerleri ile karşılaştırılır, doğru tahmin edilen (başarılı) ve doğru tahmin edilmeyen (başarısız) durumlar kaydedilir. Başarılı ve başarısız durumlar, her bir gözlenmemiş düğüm için ne kadar doğru tahminler yapıldığını gösteren ölçütlerin oluşturulmasında kullanılır. Bu ölçütlerin bazıları hata matrisi, hata oranı, kalibrasyon tablosu, logaritmik kayıp skoru, karesel (Brier) skor, küresel sonuç skoru, şaşırtma indeksi, test duyarlığı olarak verilebilir. Bu çalışmada Bayesci ağın uyumunu test etmede kayıp skorlarından yararlanılmıştır. Logaritmik kayıp skoru, 0 ile sonsuz arasında değer alır ve bu değerin sıfıra yaklaşması yüksek performansı ifade eder. Karesel kayıp skoru, 0 ile 2 arasında değer alır ve değerin sıfıra yaklaşması yüksek performansı ifade eder. Küresel kayıp skoru ise 0 ile 1 arasında değer alırken bu değerin 1'e yaklaşması yüksek performansı ifade eder [16].

3. Sayısal Örnek

Sayısal örnek için kullanılan veri kümesi, 30 adet restorana ait 6 adet değişkenden oluşmaktadır [11]. Değişkenlere ilişkin bilgiler aşağıdaki gibidir.

X_1 : Restorandaki koltuk sayısı

X_2 : Bir vardiyadaki ortalama garson sayısı

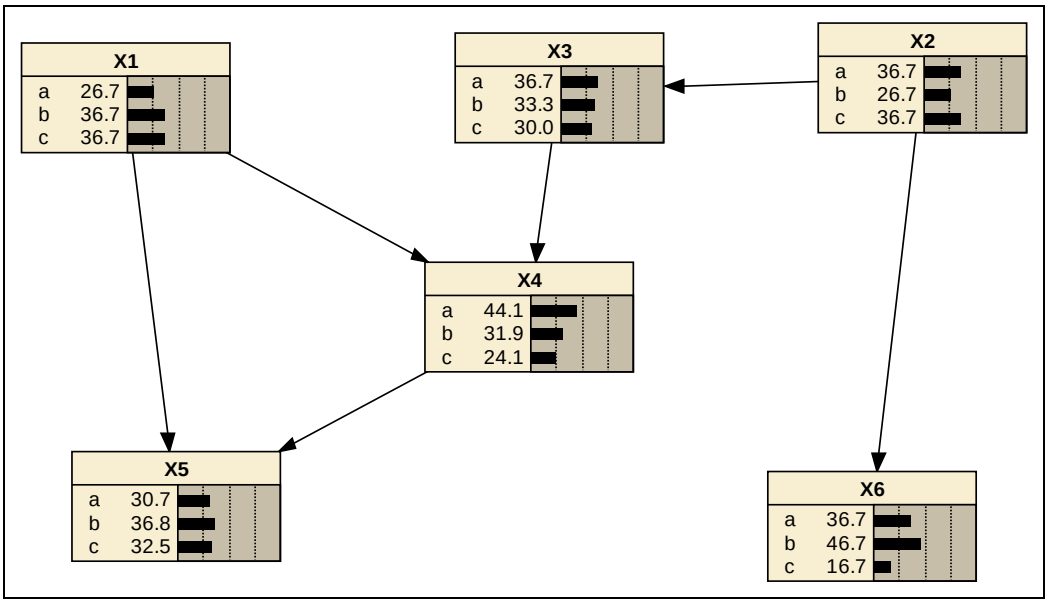
X_3 : Restorandaki ortalama personel sayısı (aşçı, şef, kasiyer vb.)

X_4 : Restoranın kapladığı alan (metrekare olarak)

X_5 : Bir günde restorana gelen müşteri sayısı

X_6 : Restorandaki bir yemeğin ortalama fiyatı

Bu veri kümesinde yer alan tüm değişkenler süreklidir. Bayesci ağ oluşturmak için tüm değişkenler üç düzeyli kategorik değişken olarak yeniden düzenlenmiştir. Bu değişkenler için bir uzman bilgisi olmadan kendi önsel bilgimize dayalı olarak oluşturduğumuz Bayesci ağ Şekil 2 ile verilmiştir.



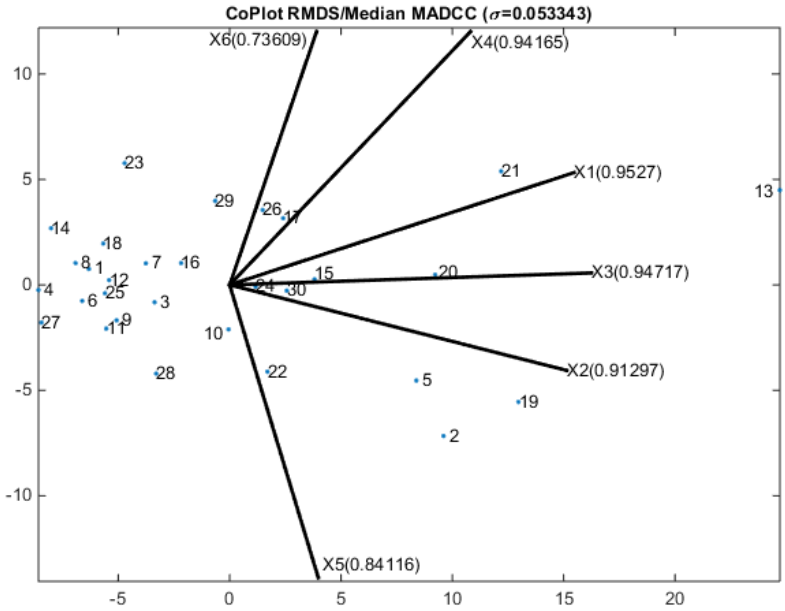
Şekil 2. Restoran verisi için oluşturulan Bayesci ağ.

Bu veri kümesinde ilgilenilen değişken "bir günde restorana gelen müşteri sayısı-X5" olsun. Bu değişken gözlenmemiş olarak seçildiğinde ağın performansını değerlendirmede kullanılan kayıp fonksiyonlarının değerleri aşağıdaki gibi elde edilmiştir.

Logaritmik kayıp skoru = 0,7808
Karesel kayıp skoru = 0,4846
Küresel kayıp skoru = 0,7139

Bayesci ağın oluşturulmasının ardından aynı veri kümesine ilişkin Robust Coplot grafiği Şekil 3 ile verilmiştir. Grafik incelendiğinde Kruskal stress değeri yaklaşık olarak 0,05 olarak bulunmuştur. Çok boyutlu ölçeklemede elde edilen grafiğin başarısının bir göstergesi olan Kruskall değeri 0,05 bulunduğu için gözlemlerin iki boyutlu düzleme indirgenmesine ilişkin başarının oldukça iyi olduğu söylenebilir [1]. Her bir değişkeni temsil eden vektörlere ilişkin korelasyon katsayıları X6 hariç oldukça yüksek bulunmuştur. X6 değişkenine ilişkin korelasyon katsayısı ise çok düşük değildir. Bu durum bize vektörlerin yerleşiminin de başarılı olduğunu ifade eder. X2 ve X3 no'lu değişkenler aynı yönde ve aralarındaki açı da küçük olduğu için restorandaki garson sayısı ile çalışan sayısına ilişkin değişkenler arasında pozitif yönde güçlü bir korelasyon olduğunu söyleyebiliriz. X6 ve X5 değişkenleri arasında ters yönde güçlü bir korelasyon vardır. Dikkat edilirse bir restorandaki yemeğin ortalama fiyatı artarsa o restorana gelecek müşteri sayısının azalması mantıklı bir çıkarım olacaktır. X6 değişkeni ile X4 değişkeni arasında pozitif yönlü bir korelasyon olduğu söylenebilir. X1, X2, X3 değişkenleri arasında aynı yönlü ilişkinin olduğu yine grafikten çıkartılacak sonuçlar arasındadır. X6 ile X2 ya da X5 ile X1 arasındaki açı yaklaşık olarak 90 derece olduğu için bu değişkenlerin ilişkisiz olduğu söylenebilir. Gözlemlerin yerleşimine bakıldığında 13 no'lu gözlemin diğer gözlemlerden daha uzakta yerleştiği dikkat

çekmektedir. Orijinal veri kümesi bu açıdan değerlendirildiğinde bu restoranın 30 restoran içinde alanı 1600 metrekare ile en büyük olan restoran olduğu ve kendisine en yakın olan alanın 1000 metrekare olduğu görülmektedir. Bu özelliği nedeniyle 13 no'lu gözlem veri kümesinin genel yerleşiminden daha uzakta yer almaktadır. 2, 5, 13, 19, 20, 21 no'lu gözlemlerin X1, X2, X3 ve X4 değişkenlerini temsil eden vektörler üzerindeki iz düşüm değerleri yüksektir. Bu sebeple bu gözlemlerin ilgili değişkenlerdeki değerlerinin de yüksek olduğu sonucu çıkartılabilir.



Şekil 3. Restoran verisi Robust Coplot grafiği.

Robust Coplot grafiğinden elde edilen ilişki yapılarından yararlanarak Şekil 2 ile verilen Bayesci ağ aşağıda verilen değişimlerin uygulanmasıyla güncellenir.

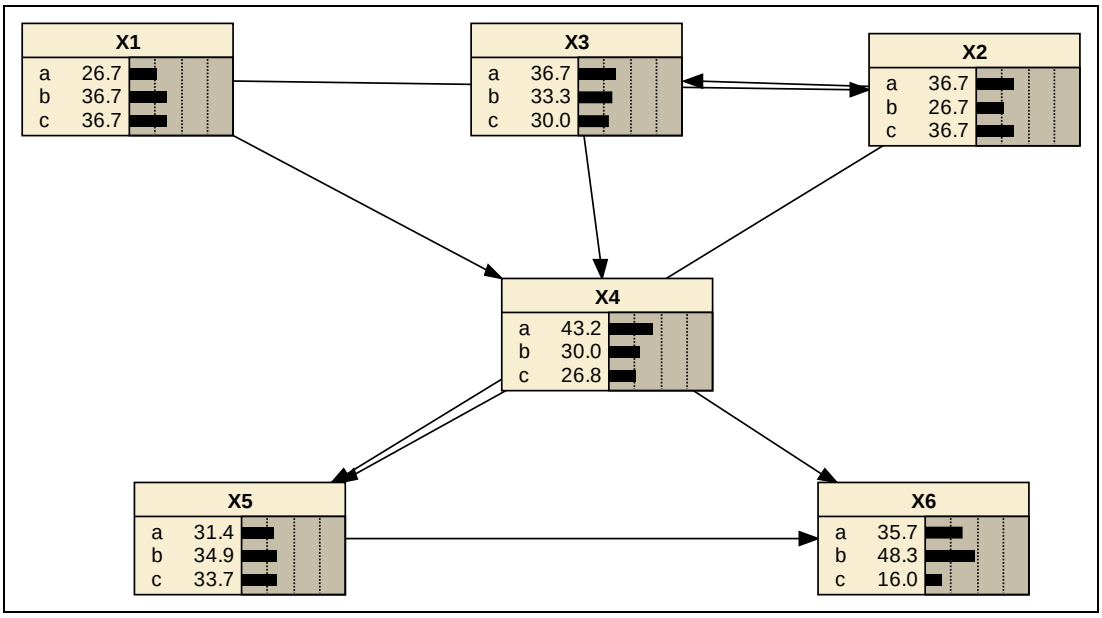
- X1→X2 bağı eklenir
- X2→X5 bağı eklenir
- X4→Y2 bağı eklenir
- X5→X6 bağı eklenir
- X1→X5 bağı çıkartılır
- X2→X6 bağı çıkartılır

Bu güncellemelerin yapılmasının ardından elde edilen yeni Bayesci ağ Şekil 4 ile verilmiştir.

Güncellenen Bayesci ağ için skor değerleri aşağıda verilmiştir.

Logaritmik kayıp skoru = 0,4337
Karesel kayıp skoru = 0,2929
Küresel kayıp skoru = 0,8344

Kayıp fonksiyonlarından da görüldüğü gibi Şekil 4 ile verilen güncellenen Bayesci ağın kendi önsel bilgimize göre oluşturduğumuz Şekil 2 ile verilen Bayesci ağdan daha iyi sonuç verdiği söylenebilir.



Şekil 4. Restoran verisi için güncellenen Bayesci ağ.

4. Sonuç ve öneriler

Bayesci ağlar genellikle uzman bilgisine dayanılarak oluşturulur. Bununla birlikte uzman bilgisine ulaşmak zor ve yüksek maliyetli bir iştir. Özellikle uzman bilgisinin olmadığı durumlarda, araştırmacıya değişkenler arasındaki ilişkileri saptamasında faydalı olacak istatistiksel yöntemlerden yararlanmak bu modelleri oluşturmayı kolaylaştıracaktır. Değişkenler arasındaki ilişki yapıları kullanılarak uzun süren öğrenme algoritmalarından kaçınmak mümkün olmakta, bu da verinin analiz sürecini ciddi ölçüde kısaltmaktadır. Bu çalışmada, çok boyutlu veri kümesindeki değişkenler ve bu değişkenlerin birbirleri ile olan ilişkileri hakkında uzman bilgisinin olmadığı durumda Bayesci ağ yapısının öğrenilmesinde kullanılmak üzere grafiksel bir yaklaşım tanıtılmıştır. Çok boyutlu veri kümesinin Robust Coplot analizi sonucu elde edilen grafiğinden yararlanarak değişkenler arasındaki ikili korelasyonlar değerlendirilmiş ve bu bilgi kullanılarak oluşturulan Bayesci ağların veriye daha uygun olduğu gözlenmiştir. Buradan yola çıkarak araştırmacının değişkenler arasındaki yapı hakkında bilgisi yok ise verinin ön incelmesini yaptıktan sonra Bayesci modellemeye geçmesinin bir avantaj olacağı, grafiksel değerlendirme sonrasında değişkenler arasındaki bağları oluştururken Robust Coplot grafiğinin faydalı olacağı bir örnek üzerinden vurgulanmıştır.

Kaynaklar

[1] Y. K. Atilgan, 2016, Robust Coplot Analysis, *Journal Communications in Statistics - Simulation and Computation*, 45 (5), 1763-1775.

[2] Ben-Gal, 2007, Bayesian Networks, *Encyclopedia of Statistics in Quality &Reliability*, F. Ruggeri, F.Faltin, R. Kenett, R. (eds), Wiley & Sons.

[3] S.G. Boettcher, C. Dethlefsen, C., 2003, deal: A package for learning Bayesian networks, *Journal of Statistical Software*, 8 (20), 1-40.

[4] D. M. Bravata, K. G. Shojania, I. Oklin, A. Raveh, 2007, CoPlot: A tool for visualizing multivariate data in medicine, *Statistics in Medicine*, 27 (12), 2234-2247.

[5] D. Chickering, D. Geiger, D. Heckerman, 1995, Learning Bayesian networks: Search methods and experimental results, *Proceedings of Fifth Conference on Artificial Intelligence and Statistics, Ft. Lauderdale*, FL, 112–128.

[6] H. Demirhan, Y. K. Atilgan, 2015, New horizontal global solar radiation estimation models for Turkey based on robust coplot supported genetic programming technique, *Energy Conversion and Management*, 106, 1013-1023.
[7] Y. Dong-Peng, L. Jin-Lin, 2008, Research on personal credit evaluation model based on bayesian network and association rules, *Knowledge Discovery and Data Mining, 2008. WKDD 2008. First International Workshop on*, 457-460.
[8] D. Ersel, 2012, *An Original Combined Interestingness Measure in Association Analysis*, Unpublished PhD Thesis, Hacettepe University Institute of Graduate Studies in Science, Ankara, Türkiye.
[9] P. A. Ferero, G. B. Giannakis, 2011, Robust multi-dimensional scaling via outlier sparsity control, *In: 45th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA. pp. 1183–1187.
[10] Y. Goldreich, A. Raveh, 1993, COPLOT Display Technique as an Aid to Climatic Classification, *Geographical Analysis*, 25 (4), 337-353.
[11] Y. Hadad, L. Friedman, M. Z. Hanani, 2007, Measuring Efficiency Of Restaurants Using the Data Envelopment Analysis Methodology, *Applied Statistics Computer Modelling and New Technologies*, 11 (4), 25-35.
[12] D. Heckerman, 1995, Bayesian networks for data mining, *Data Mining and Knowledge Discovery*, 79-119.
[13] F.V. Jensen, 2001, *Bayesian Networks and Decision Graphs*, Springer-Verlag, New York, 268p.
[14] G. Lipshitz, A. Raveh, 1994, Application of the Co-plot method in the study of socioeconomic differences among cities: A basis for a differential development policy, *Urban Studies*, 31, 123-135.
[15] B. Mahlberg, A. Raveh, 2012, Co-plot: A useful tool to detect outliers in DEA, *Available at SSRN: http://ssrn.com/abstract=1999370 or http://dx.doi.org/10.2139/ssrn.1999370*.
[16] Netica, 2016, Testing Nets with Cases, *Tutorial on Bayesian Networks with Netica*, http://www.norsys.com/tutorials/netica/secD/tut_D2.htm.
[17] A. Raveh, 2000, The Greek banking system: Reanalysis of performance, *European Journal of Operational Research*, 120, 525-534.
[18] G. Shevlyakov, P. Smirnov, 2011, Robust estimation of the correlation coefficient: An attempt of survey, *Austrian Journal of Statistics*, 40, 147-156.
[19] D. Talby, D. G. Feitelson, A. Raveh, 1999, Comparing logs and models of parallel workloads using the co-plot method, *Lecture Notes in Computer Science*, 1659, 43-66.
[20] P. C. C. Wang, 1978, *Graphical Representation of Multivariate Data*, Akademic Press, New York.