IAAOJ, Scientific Science, 2015, 3(1), 19-24

SKORLAMA ÖLÇÜTLERİNİN BAYESCİ YAPI ÖĞRENME ALGORİTMALARI ÜZERİNDEKİ ETKİLERİNİN İNCELENMESİ

Emre DÜNDER[1], Mehmet Ali CENGİZ[1], Serpil GÜMÜŞTEKİN[1]
[1]Ondokuz Mayıs Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, Samsun, Türkiye

E-posta: emre.dunder@omu.edu.tr

ÖZET

Yapı öğrenme algoritmalarında skorlama ölçütleri, oluşturulan Bayesci ağ yapısının performansını doğrudan etkilemektedir. Bu çalışmada farklı skorlama ölçütlerinin skor tabanlı ve karma algoritmalar kullanılarak oluşturulan Bayesci ağ yapısı üzerindeki etkisi araştırılmıştır. Alarm veri seti kullanılarak dört farklı skorlama ölçütü üç farklı Bayesci yapı öğrenme algoritması uygulanarak karşılaştırılmıştır. Elde edilen sonuçlara göre Akaike bilgi kriteri (AIC) ve Bayesci bilgi kriteri (BIC) skorlama ölçütleri ile oluşturulan yapılar Bayesci Dirichlet eşdeğerlilik (BDe) ve K2 skorlama ölçütlerine göre daha iyi sonuç vermiştir.

Anahtar Kelimeler: Bayesci Ağlar, Yapı Öğrenme, Skorlama Ölçütleri, Skor Tabanlı Algoritmalar, Karma Algoritmalar

EXAMINING THE AFFECTS OF SCORING METRICS ON BAYESIAN NETWORKS STRUCTURE LEARNING

ABSTRACT

The scoring metrics directly affect the performance of the constructed Bayesian network structure. In this study we investigated the affects of different scoring metrics on Bayesian network structure using score based and hybrid algorithms. By using Alarm data set we compared four different scoring metrics with applying three Bayesian network structure learning algorithms. According to obtained results, the structures which constructed with Akaike information criteria (AIC) and Bayesian information criteria (BIC) gives better results according to Bayesian Dirichlet equalivance and K2 scoring metrics.

Keywords: Bayesian networks, Structure learning, Scoring metrics, Scoring Criteria, Score-Based Algorithms, Mixed Algorithms

1. Giriş

Bayesci ağlar, çok değişkenli bir veri yapısı içerisinde bulunan koşullu bağımsızlık ve nedensellik ilişkilerini grafikler aracılığıyla betimleyen grafiksel modellerdir. Bir Bayesci ağ temel olarak bir yönlü döngüsüz grafik ve koşullu olasılık tablosundan oluşur. Grafiksel yapı değişkenler arasındaki bağımlılık ilişkilerini belirten yönlü döngüsüz grafik şeklinde gösterilir. Grafiğin içerisinde bulunan her düğüm bir değişkene karşılık gelir. Düğümler arası yönlü kenarlar nedensellik ilişkilerini gösterir. Koşullu olasılık tablosu Bayesci ağın içerisinde bulunan her düğüm için koşullu olasılık değerlerini içerir. Düğümlerin her koşullu olasılık değeri parametre olarak adlandırılır (Neapolitan, 2003).

Bayesci ağlar, değişkenler arasındaki nedensellik, bağımsızlık ve koşullu bağımsızlık ilişkilerini yorumlamaya elverişli olduğundan birçok farklı bilim dalında kullanılmaktadır. Bayesci ağlara ilişkin ilk çalışma (Pearl, 1985) tarafından gerçekleştirilmiştir. Son yıllarda tıp, biyoloji, fizik vb alanda Bayesci ağlar kullanılmıştır. Bayesci ağlar akciğer kanseri teşhisi (Ramirez vd. , 2007) , Bayesci gen ağı tahmini (Dawy vd. , 2011), kök sistemi (Zhong vd., 2011) gibi farklı konular için uygulanmıştır.

Bir Bayesci ağı oluştururken yapı ve parametre olmak üzere iki temel kısım elde edilmelidir. Bayesci ağın yönlü döngüsüz grafiği yapı, koşullu olasılık değerleri parametre olarak adlandırılır (Neapolitan, 2003). Bir Bayesci ağı oluşturma sürecine öğrenme denir. Bayesci ağlarda öğrenme yapı ve parametre öğrenme olmak üzere ikiye ayrılır. Yapı öğrenme, çok değişkenli bir veri setinden yararlanılarak değişkenler arası ilişkileri gösteren yönlü döngüsüz grafik (y.d.g) oluşturma işlemidir. Parametre öğrenme, ebeveynleri bilinen değişkenlerin koşullu olasılık değerlerinin tahmin edilmesidir. Bu çalışmada yapı öğrenme algoritmaları üzerinde durulacaktır. Çalışmanın ilk bölümünde Bayesci yapı öğrenme algoritmaları ve skorlama ölçütleri tanıtılacaktır. Çalışmanın ikinci bölümünde dört farklı skorlama ölçütü ve üç farklı Bayesci yapı algoritması kullanılarak oluşturulan Bayesci ağlar karşılaştırılmıştır. Bayesci ağları karşılaştırmak üzere Asya (Lauritzen ve Spiegelhalter, 1988) ve Sigortacılık (Binder vd., 1997) veri setleri kullanılmıştır. Farklı örnek birimleri için tepe tırmanma, tabu araştırma ve en az en çok tepe tırmanma algoritmaları, K2, AIC, BIC ve BDe skorları kullanılarak elde edilen ağlar için BDe skorları karşılaştırılmıştır. Her örnek birim için 100 iterasyon kullanılarak skorların ortalama değerleri değerlendirilmiştir. Tahmin edilen Bayesci ağ yapılarının kalitesini değerlendirmek üzere gerçek ağ yapıları ile elde edilen Bayesci ağlar arasındaki farklılıkları inceleme üzere yapısal ham uzaklıkları elde edilmiştir.

2. Materyal ve Metot

Bu bölümde skor tabanlı ve karma algoritmalar üzerinde durulmuştur. Skor tabanlı algoritmalardan tepe tırmanma ve tabu araştırma algoritmaları tanıtılacaktır. Karma tabanlı algoritmalardan en az en çok tepe tırmanma algoritması üzerinde durulmuştur.

2.1. Skor Tabanlı Algoritmalar

Skor tabanlı algoritmalar bir arama yöntemi ve skorlama ölçütünden yararlanılarak bir Bayesci ağ yapısı oluşturur. Skor tabanlı algoritmalarda amaç bir arama yöntemi kullanılarak skorlama ölçütünün değerini en yüksek şekilde oluşturacak bir yapı elde etmektir (Gamez vd., 2011). Bir diğer ifade ile skor tabanlı algoritmalar mümkün olan tüm ağlar arasından en yüksek skor değerine sahip olan Bayesci ağı bulmayı amaçlar (Jensen ve Nielsen, 2007). Literatürde sıkça kullanılan yöntemlerden biri de K2 skor ölçütü ile aç gözlü arama yöntemini birleştiren K2 algoritmasıdır (Cooper ve Herkovits, 1992). Skor tabanlı algoritmalar içinde en çok kullanılan algoritmalardan biri aç gözlü arama (greedy search) algoritmasıdır. Bu algoritmalar uygulanırken tepe tırmanma, tabu araştırma (Bouckaert, 1995), açgözlü eşdeğer arama (Chickering, 2002a) gibi yöntemler kullanılır (Scutari, 2011).

Bu yöntemlerde Bayesci ağ yapısını oluştururken üç temel işlem gerçekleştirilir: kenar ekleme, kenar silme ve kenarı tersine çevirme. Kenar ekleme, iki düğüm arasına yönlü bir kenar çizimini öngörür. Kenar silme, iki düğüm arasında bulunan yönlü kenarı silme işlemidir. Kenarı tersine çevirme, iki düğüm arasındaki kenarın yönünü ve dolayısıyla nedensellik ilişkisini değiştirir. Gerçekleştirilen her işlem için bir skorlama ölçütü aracılığıyla bir skor değeri hesaplanır. Elde edilen skor değerine göre düğümler arası uygulanacak işleme karar verilir. Yerel arama algoritmalarında, skor değerini en çok artıracak olan kenar işlemi uygulanır. Bu işlemler, skorlama ölçütünde bir artış olmadığı duruma kadar devam eder.

2.2. Karma Tabanlı Algoritmalar

Karma algoritmalar kısıtlama tabanlı algoritmalarda kullanılan koşullu bağımsızlık testlerinden ve skor tabanlı algoritmaların skor ölçütlerinden yararlanarak Bayesci ağ yapısını oluşturmaktadır. Bu algoritmaların temelindeki mantık kısıtlama tabanlı algoritmalardan yararlanarak bir yönsüz grafik oluşturmak ve ardından skor tabanlı algoritmaları uygulayarak kenarların yönlerini belirlemektir.

Literatürde uygulanan ilk karma algoritma CB algoritmasıdır (Singh ve Valtorta, 1993). Bu yöntem SGS (Spirtes vd., 1990) ve K2 algoritmalarının birleşiminden oluşmaktadır. Kesim kümeleri kavramını ve d-ayırma kurallarına dayalı olarak Benedict (Acid ve Campos, 2001) algoritması önerilmiştir. Sahte aday (SC) algoritması (Friedman vd. , 1999) karma algoritmalar arasında en sık kullanılan yöntemlerden birisidir. Bu algoritmanın temel amacı, her düğüm için en fazla k adet ebeveyne sahip bir aday ebeveynler kümesi oluşturarak arama uzayını kısıtlamaktır. Kısıtlandırılmış ebeveynler kümesi üzerinden tepe tırmanma algoritması uygulanarak kenarlar yönlendirilerek Bayesci ağ oluşturulur. En az en çok tepe tırmanma (MMHC) algoritması (Tsamardinos vd. , 2006) sahte aday

algoritmasının geliştirilmiş şeklidir. Bu algoritmada sahte aday algoritmasından farklı olarak her düğüm için gerçekleştirilen kısıtlandırma işlemi en az en çok ebeveynler ve çocuklar (MMPC) algoritması kullanılarak oluşturulur. En az en çok ebeveynler ve çocuklar algoritması, sahte aday algoritması tarafından oluşturulan aday kümesi skor ölçütü ve gerçek ağa olan yapısal yakınlığı bakımından daha doğru sonuçlar vermektedir. En az en çok tepe tırmanma algoritmasının genelleştirilmiş şekli iki aşamalı kısıtlandırılmış en yüksek değere ulaştırma algoritmasıdır. Bu algoritma ile kısıtlandırma ve kenar yönlendirme aşamalarında farklı yöntemler kullanılarak Bayesci ağ oluşturulmaktadır.

2.3. Skorlama Ölçütleri

Literatürde en sık kullanılan skorlama ölçütleri kesikli ve sürekli veriler için yukarıdaki tabloda gösterilmektedir. Bu çalışmada kesikli veriler kullanılarak Bayesci ağ oluşturulacaktır. Kesikli veriler için kullanılan BIC (Schwarz, 1978) ve AIC (Akaike, 1974), bilgi tabanlı skorlama ölçütleridir. Bu ölçütler verinin olabilirliğinin yanı sıra modelin karmaşıklığını ortaya koymak üzere bir ceza terimi kullanır. Ceza terimi ağ yapısı için tahmin edilecek olasılık sayısına bağlıdır (Bouckaert, 1995). BDe skoru (Heckerman vd. , 1995) bir veri seti verildiğinde Bayesci ağın sonsal olasılığını hesaplamaya dayanır. BDe skoru hesaplanırken önsel olarak düzgün dağılım kullanılır. Ayrıca BDe skorlama ölçütü, skor eşitliği varsayımını sağlamaktadır. Bu varsayıma göre aynı iskelete ve aynı koşullu olasılık yapılarına sahip olan Bayesci ağlar aynı skor değerine sahiptir (Verma ve Pearl, 1991).

	Veri Tipi	
	Kesikli Veriler	**Sürekli Veriler**
Skorlama Ölçütü	K2 Skoru Bayesci Bilgi Ölçütü Akaike Bilgi Ölçütü BDe (Bayesci Dirichlet Eşdeğerlilik) Skoru	Bayesci Gausyen Eşdeğerlilik Skoru Akaike Bilgi Ölçütü Bayesci Bilgi Ölçütü

Bu çalışmada skor ölçütlerinin skor tabanlı ve karma yapı öğrenme algoritmaları üzerindeki etkileri araştırılacaktır. Skor tabanlı yöntemler için tepe tırmanma ve tabu araştırma algoritmaları kullanılacaktır. Karma algoritmalar için en az en çok tepe tırmanma ve iki aşamalı kısıtlandırılmış en yüksek değere ulaştırma algoritmaları kullanılacaktır. Uygulanacak tüm yapı öğrenme algoritmaları için K2, BIC, AIC ve BDe skorları ile Bayesci ağlar oluşturulacak ve farklı skorlara göre hem skor tabanlı, hem de karma algoritmaların tahmin performansı karşılaştırılacaktır. Tahmin performansları literatürde sıkça kullanılan BDe skoruna göre karşılaştırılacaktır.

3. Uygulama

Bu çalışmada literatürde sıkça kullanılan Asya ağı ve Sigortacılık ağı kullanılmıştır. Her iki ağ için de 100, 500 ve 1000' er birimlik örneklem seçilerek 100 iterasyon sonucunda elde edilen skor ölçütlerinin aritmetik ortalamaları hesaplanmıştır. Skor ölçütlerinin ortalama değerlerine göre skor ölçütlerinin yapı öğrenme algoritmaları üzerindeki etkileri karşılaştırılmıştır. Ayrıca farklı skor ölçütleri için skor tabanlı ve karma algoritmaları ile kurulan ağların performansları incelenmiştir. Bayesci ağların yapılarını değerlendirirken performans ölçütü olarak literatürde sıkça kullanılan BDe skoru ve yapısal ham uzaklık değerleri kullanılmıştır. Asya ve Sigortacılık verileri için farklı sayıdaki örneklemler çekilerek tepe tırmanma, tabu araştırma, en az en çok tepe tırmanma ve iki aşamalı kısıtlandırılmış en yüksek değere ulaştırma algoritmaları K2, AIC, BIC ve BDe skor fonksiyonları kullanılarak ayrı ayrı oluşturulmuştur. Elde edilen Bayesci ağlar ile gerçek ağ yapıları arasındaki farklılıklar yapısal ham uzaklık değerleri ile karşılaştırılmıştır.

Tablo 1: Tepe tırmanma algoritması için BDe skor değerleri

	TEPE TIRMANMA					
	N=100		N=500		N=1000	
	ASYA	SİGORTA	ASYA	SİGORTA	ASYA	SİGORTA
K2	-11.453,97	-307.194,90	-11.135,14	-283.998,60	-11.112,96	-274.374,90
AIC	**-11.410,62**	-306.769,00	**-11.116,07**	-283.697,80	**-11.097,73**	-273.885,10
BIC	-11.423,10	**-306.712,00**	-11.119,14	**-283.677,40**	-11.097,88	**-274.009,10**
BDe	-11.492,75	-307.113,70	-11.176,43	-283.885,10	-11.150,67	-273.995,30

Tablo 2. Tabu araştırma algoritması için BDe skor değerleri

	TABU ARAŞTIRMA					
	N=100		N=500		N=1000	
	ASYA	SİGORTA	ASYA	SİGORTA	ASYA	SİGORTA
K2	-11.447,44	-305.990,90	-11.134,92	-282.487,70	-11.113,01	-273.051,70
AIC	-11.413,66	**-305.404,20**	-11.116,26	**-282.094,50**	**-11.097,57**	**-272.685,70**
BIC	**-11.405,69**	-305.417,00	**-11.113,27**	-282.135,50	-11.097,78	-272.746,00
BDe	-11.488,44	-305.821,20	-11.174,13	-282.278,30	-11.151,70	-272.822,80

Tablo 3. En az en çok tepe tırmanma algoritması için BDe skor değerleri

	EN AZ EN ÇOK TEPE TIRMANMA					
	N=100		N=500		N=1000	
	ASYA	SİGORTA	ASYA	SİGORTA	ASYA	SİGORTA
K2	-12.353,35	-340.321,40	-12.119,58	-306.397,40	-12.094,58	-294.691,90
AIC	**-12.316,91**	**-339.481,90**	**-12.096,73**	-306.049,10	-12.080,48	**-294.066,80**
BIC	-12.318,80	-339.854,00	-12.096,79	**-305.995,50**	**-12.080,23**	-294.080,20
BDe	-12.391,83	-340.193,90	-12.152,43	-306.375,60	-12.130,51	-294.215,20

Tablo 4. Tepe tırmanma algoritması için SHD uzaklık değerleri

	SHD HC					
	N=100		N=500		N=1000	
	ASYA	SİGORTA	ASYA	SİGORTA	ASYA	SİGORTA
K2	8,449	55,876	6,501	42,733	6,227	**40,000**
AIC	6,501	52,722	3,894	44,323	**3,498**	40,713
BIC	**3,000**	**48,000**	**3,000**	**41,000**	4,000	41,000
BDe	10,481	65,686	8,465	52,267	7,247	49,785

Tablo 5. Tabu araştırma algoritması için SHD uzaklık değerleri

	SHD TABU					
	N=100		N=500		N=1000	
	ASYA	SİGORTA	ASYA	SİGORTA	ASYA	SİGORTA
K2	8,670	56,412	6,517	43,495	5,790	40,320
AIC	**6,908**	**51,894**	4,126	42,058	**3,789**	**35,468**
BIC	9,000	54,000	**4,000**	**42,000**	5,000	37,000
BDe	11,569	66,422	9,325	50,555	7,968	47,735

Tablo 6. En az en çok tepe tırmanma algoritması için SHD uzaklık değerleri

	SHD MMHC					
	N=100		N=500		N=1000	
	ASYA	SİGORTA	ASYA	SİGORTA	ASYA	SİGORTA
K2	6,675	55,740	5,403	51,757	4,969	50,613
AIC	6,370	55,825	**5,013**	50,750	4,623	49,575
BIC	9,000	**55,000**	8,000	**50,000**	**4,000**	**46,000**
BDe	**6,366**	54,979	5,088	50,901	4,694	49,132

Tablo 1, 2 ve 3' te Asya ve Sigortacılık verileri için farklı örneklem düzeylerinde dört farklı algoritma ve dört farklı skor ölçütü kullanılarak elde edilen ağlara ilişkin BDe skor değerleri verilmiştir. Buna göre hem skor tabanlı hem de karma algoritmalar için AIC ve BIC skor ölçütleri ile oluşturulan Bayesci ağların skor değerlerinin daha yüksek olduğu görülmektedir. Ayrıca yapı öğrenme algoritmalarını karşılaştırdığımızda skor tabanlı algoritmalar ile elde edilen ağlarda karma algoritmalar ile elde edilenlere göre daha yüksek BDe skorları hesaplanmıştır.

Tablo 4, 5 ve 6' da dört farklı skor ölçütü ve yapı öğrenme algoritması kullanılarak elde edilen Bayesci ağların gerçek ağlar arasındaki farklılıkları ölçmek için elde edilen yapısal ham uzaklık değerleri grafiksel biçimde gösterilmektedir. Sonuçlara bakıldığında genel olarak K2 ve BDe skor ölçütleri ile kurulan ağların AIC ve BIC skor ölçütleri ile kurulan ağlara göre daha yüksek yapısal uzaklık değerlerine sahip oldukları görülmektedir. Buna göre AIC ve BIC skor ölçütleri ile kurulan ağların K2 ve BDe skor ölçütleri ile kurulan ağlara göre daha doğru bir Bayesci ağ yapısı tahmin etmektedir.

4. Sonuç ve Değerlendirme

Bu çalışmada farklı skorlama ölçütlerinin Bayesci ağ yapısının kalitesi üzerinde etkisi incelenmiştir. Ele edilen sonuçlara göre AIC ve BIC skorlama ölçütleri kullanılarak oluşturulan yapıların K2 ve BDe skorlarına göre daha yüksek skor değerine sahip olduğu belirlenmiştir. Buna göre Bayesci ağ yapısı için kullanılan skorlama ölçütlerinin son derece önemli olduğu görülmektedir. AIC ve BIC skorlama ölçütlerinin yapı öğrenme algoritmaları için daha elverişli sonuçlar verdiği sonucuna varılmaktadır. AIC ve BIC skorlama ölçütleri içerisinde bulunan cezalandırma terimlerinin Bayesci ağ yapısının tahmin performansını artırdığı söylenebilir.

Kaynaklar

Acid, S., de Campos, L., M. (2003). Searching for Bayesian network structures in the space of restricted acyclic partially directed graphs. Journal of Artificial Intelligence Research, 445–490.

Akaike, H., (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 716-723.

Binder, J., Koller, D., Russell, S., Kanazawa., K., (1997). Adaptive Probabilistic Networks with Hidden Variables. Machine Learning, 29, 213-244.

Bouckaert, R. R. (1995). Bayesian Belief Networks: from Construction to Inference. PhD thesis, Utrecht University, The Netherlands.

Chickering, D., M.(2002). Optimal Structure Identification with Greedy Search. Journal of Machine Learning Resesearch, 3,507-554.

Cooper, G., F., Herskovits, E. (1992). A Bayesian Method for Constructing Bayesian Belief Networks from Databases. Machine Learning, 9, 309–347.

Dawy, Z., Yaacoub, E., Nassar, M., Abdallah, R., Zeineddine, H., A. (2011). A multiorganism based method for Bayesian gene network estimation. BioSystems, 103, 425–434.

Friedman, N., Nachman, I., Peéer D. (1999). Learning Bayesian Network Structure fromMassive Datasets: The "Sparse Candidate" Algorithm. Proc. Fifteenth Conf. on Uncertainty in Artificial Intelligence (UAI)

Gamez., J., A., Mateo., J., L., Puerta, J., M. (2011). Learning Bayesian Networks by hill-Climbing: efficient methods based on progressive restriction of the neighborhood. Data Mining Knowledge, 22, 106-148.4

Heckerman, D., Geiger, D., Chickering, D., M. (1995). Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Machine Learning, 20, 197-243.

Jensen, F., V., Nielsen, T., D. (2007) Bayesian Networks and Decision Graphs. Information Science and Statistics.

Lauritzen, S., Spiegelhalter, D. (1988). Local Computation with Probabilities on Graphical Structures and their Application to Expert Systems. Journal of the Royal Statistical Society, 50, 157-224.

Neapolitan, R., (2003). Learning Bayesian Networks. Prentice Hall Series in Artificial Intelligence

Pearl, J. (1985). Bayesian Networks: A Model of Self-Activated Memory for Evidental Reasoning. UCLA Technical Report CSD-850017. Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA. pp. 329–334.

Ramírez,, N., C., Mesaa, H. G. A., Calvet, H., C., Fernández, L., A., N., Martínez, R., E., B. (2007). Diagnosis of breast cancer using Bayesian networks:A case study. Computers in Biology and Medicine, 37, 1553 – 1564.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 461-464.

Scutari, M. (2011). Measures of Variability for Graphical Models. PhD thesis, Universita degli Studi di Padova, Dipartimento di Scienze Statistiche.

Singh, M., & Valtorta, M. (1993). An algorithm for the construction of Bayesian network structures from data. In 9th Conference on Uncertainty in Artificial Intelligence, pp. 259–265.

Spirtes, P., Glymour, C., & Scheines, R. (1990). Causality from probability. Evolving knowledge in the natural and behavioral science, 181–199.

Tsamardinos, I., Brown L., E., Aliferis C., F. (2006). The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. Machine Learning, 65, 31-78.

Verma, T., S., Pearl, J. (1991). Equivalence and Synthesis of Causal Models. Uncertainty in Artifcial Intelligence, 6, 255-268.

Zhong, J., Huang, R., S., Lee, S., C. (2011). A program for the Bayesian Neural Network in the ROOT framework. Computer Physics Communications, 182, 2655–2660