

RASTLANTISAL ÖBEK ÇİZGELER İÇİN BAYESÇİ MODEL SEÇİMİ

BAYESIAN MODEL SELECTION OF STOCHASTIC BLOCKMODELS

FOR RANDOM GRAPHS

Barış Kurt, A. Taylan Cemgil

Algısal Zeka Lab.
Bilgisayar Mühendisliği Bölümü
Boğaziçi Üniversitesi
{baris.kurt,taylan.cemgil}@boun.edu.tr

ÖZETÇE

Gözlemlenen bir veriyi elimizdeki olasılık modellerinin hangisinin daha iyi açıkladığı problemini çözmenin bir yolu Bayesçi model seçimidir. Bu makalede en basit rastgele çizge modelleri olan Erdős-Rényi ve rastlantısal öbek modelleri için Bayesçi model seçimi uyguladık. Bir çizgenin bitişiklik matrisi verildiğinde, değişik modelleri kesin hesaplama, varyasyonel yöntem ve Monte Carlo yöntemi kullanarak hesaplanan marjinal olabilirlik değerleri üzerinden karşılaştırdık. Sentetik verilerde bu çıkarım yöntemlerini model mertebesini kestirme başarısı açısından karşılaştırdık. Çıkarım yöntemlerinin birbirlerine yakın sonuçlar verdiğini ve Monte Carlo methodunun kesin çözüme daha çok yaklaştığını gözlemledik.

ABSTRACT

A way of solving the problem of which model explains an observation better is Bayesian model selection. In this paper, we applied Bayesian model selection for the simplest graph models: the Erdős-Rényi and Stochastic Blockmodel graphs. Given the adjacency matrix of a graph, we compared its' marginal likelihood under different models using direct computation, variational methods and Monte Carlo methods. We compared the success of the methods according to their ability to estimate the correct model order. Both methods gave qualitatively similar results but the Monte Carlo method estimated the true Marginal likelihood more accurately.

1. Giriş

Çizgeler günümüzde bir çok farklı alandaki verilerin temsilinde kullanılan önemli matematiksel nesnelerdir. Biyolojik ağlar, protein etkileşim ağları, sosyal iletişim ağları gibi yapılar çizgeler vasıtasıyla temsil edilmektedir [1]. Bu verilerin modellenmesi, ve verilerdeki zaman içerisindeki değişimlerin takip edilmesinin yolu, verileri temsil eden çizgelerin matematiksel modellenmeleriyle mümkündür.

Model seçimi bir çizgeyi hangi model ile ifade etmemizin daha uygun olacağı sorusuna yanıt arar. İstatistiksel testlerin model seçiminde nasıl kullanıldığını Olding [2] özetlemiştir. Biz de bu çalışmamızda Bayesçi bir bakış açısıyla, verilen bir çizgenin farklı modeller için marjinal olabilirlik değerlerini

karşılaştırarak uygun modelin nasıl seçildiğini gösterdik. Çalışmamızda iki temel çizge modeli olan Erdős-Rényi ve rastlantısal öbek modellerini yönlendirilmiş çizgeler üzerinde uyguladık. Uyguladığımız yöntemler yönlendirilmemiş çizgeler için de kolayca genelleştirilebilir.

2. Rastgele Çizgeler

Bir çizge boğumlar ve bağlardan oluşur: $G = \{\mathcal{V}, \mathcal{E}\}$. Bağlar iki boğum arasında bir ilişki tanımlar ve bu ilişki reel bir sayı ile ifade edilebilir. Bizim inceleyeceğimiz modellerde bu bağlar ikili sayılarla ifade edilecek, yani herhangi iki boğum arasında bir bağ olma durumu 1, ve olmama durumu 0. Bir çizge matematiksel olarak en kolay bitişiklik matrisi ile ifade edilir. Eğer boğum sayımız N ise, $N \times N$ boyunda bir A bitişiklik matrisi kullanmamız gerekir, öyle ki A_{ij} değeri bize i ve j boğumları arasındaki bağ değerini versin.

2.1. Erdős-Rényi Çizgeleri

En basit rastgele çizge modeli olan Erdős-Rényi çizge modeli tek bir b parametresi ile ifade edilir. Bu parametre herhangi iki boğum arasında bir bağ olup olmama ihtimalini belirten bir Bernoulli parametresidir. Erdős-Rényi modelinin yönlendirilmiş çizgeler için üretici modelini yazacak olursak:

$$A_{ij} \sim \mathcal{BE}(A_{ij}; b) \quad \forall i, j \in \{1, 2, \dots, N\} \quad (1)$$

Eğer modelin yeterli istatistiklerini tanımlayacak olursak,

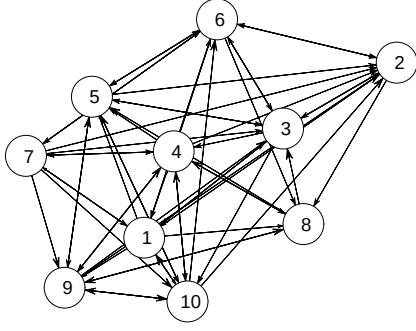
$$c = \sum_{i,j}^N A_{ij} \quad (\text{toplam bağ sayısı}) \quad (2)$$

$$n = \sum_{i,j}^N 1 \quad (\text{olabilecek tüm bağların sayısı}) \quad (3)$$

modelin b parametresi verildiğindeki koşullu olasılığını bir ikitermimli dağılım olarak şu şekilde hesaplarız:

$$p(A|b) = b^c (1-b)^{n-c} \quad (4)$$

Erdős-Rényi modelinden rastgele üretilmiş bir çizge Şekil 1'de gösterilmiştir.



Şekil 1: Erdős-Rényi modelinden oluşturulmuş bir çizge. Tüm boğumlar arası bağ olma ihtimali eşit, öbekleşme yok.

2.2. Rastlantısal Öbek Çizgeler

Bir diğer çizge modeli olan rastlantısal öbek modelinde her boğum bir kategoriye aittir ve boğumlar arası bağ oluşum olasılığı boğumların kategorilerine göre belirlenir. Eğer modelimizde K kadar kategori var ise, modeli $K \times K$ boyutundaki bir B matrisi ile ifade ederiz, öyle ki B_{rs} bize r kategorisindeki bir boğum ile s kategorisindeki bir boğum arasındaki bağ olasılığını versin. B_{rs} burada Erdős-Rényi modelinde olduğu gibi bir Bernoulli parametresidir. Kategori atamalarını $N \times K$ boyutunda bir C matrisi ile ifade edelim ve $C_{i,r}$, i boğumunun r kategorisine ait olmasını gösterebilir. Üretici modelimizi şu şekilde yazabiliriz:

$$A_{ij} \sim \mathcal{BE}(A_{ij}; B_{rs}), \text{ eğer } C_{i,r} = C_{j,s} = 1 \text{ ise} \quad (5)$$

Bir rastlantısal öbek çizgenin A bitişiklik matrisi, B parametreleri ve kategori atamaları verildiğinde, koşullu olasılığı şu şekilde hesaplarız:

$$P(A|B, C) = \prod_{i,j} \prod_{r,s} (B_{rs}^{A_{ij}} (1 - B_{rs})^{(1-A_{ij})})^{C_{i,r} C_{j,s}} \quad (6)$$

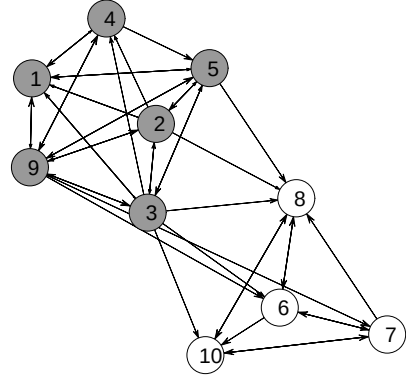
Bu modelde her kategori ikilisi r, s bir öbek oluşturmaktadır ve öbekler başlı başlarına birer B_{rs} parametresiyle ifade edilen bir Erdős-Rényi modelidir. Rastlantısal öbek çizgelerin yeterli istatistikleri de her öbeğin kendi içindeki Erdős-Rényi yeterli istatistikleridir. Bunlara c_{rs} ve n_{rs} diyecek olursak, koşullu olasılığı şu şekilde ifade edebiliriz:

$$P(A|B, C) = \prod_{r,s} B_{rs}^{c_{rs}} (1 - B_{rs})^{n_{rs} - c_{rs}} \quad (7)$$

Rastlantısal öbek modelinden rastgele üretilmiş bir çizge Şekil 2'de gösterilmiştir. Şekilde anlaşılabilirlik açısından boğumların kategorileri renklerle kodlanmış olsa da, gerçekte veri gözlemlenirken kategoriler gizli kalır.

3. Bayeşçi Model Seçimi

Bayeşçi model seçimini uygulamak için verilen bir çizgenin modeller altındaki marjinal olabilirlik değerini kıyaslamamız



Şekil 2: Rastlantısal öbek modelinden oluşturulmuş, 2 öbekli bir çizge. Öbeklerdeki boğumların birbirleriyle ve diğer öbekteki boğumlarla bağ yapma olasılıkları farklıdır, ve öbekler gözlemlenmemektedir.

gerekir. Hesap kolaylığı açısından bu değerlerin logaritmalarını kıyaslayacağız. Erdős-Rényi modeli için marjinal olabilirliğini şu şekilde hesaplayabiliriz:

$$p(A) = \int_b p(b) p(A|b) db \quad (8)$$

Bu olabilirliği hesaplamak için ise b için bir önsel olasılık tanımlamamız gerekir. Bunun için ikiterimli dağılımın eşlenik önseli olan Beta dağılımını kullandığımızda aşağıdaki denklemleri elde ediyoruz:

$$p(b) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} b^{\alpha-1} (1-b)^{\beta-1} \quad (9)$$

$$\log p(A) = \log \int_0^1 \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} b^{\alpha+c-1} (1-b)^{\beta+n-c-1} db \quad (10)$$

$$\begin{aligned} &= \log \Gamma(\alpha + \beta) + \log \Gamma(\alpha + c) \\ &\quad + \log \Gamma(\beta + n - c) - \log \Gamma(\alpha) \\ &\quad - \log \Gamma(\beta) - \log \Gamma(\alpha + \beta + n) \end{aligned} \quad (11)$$

Buradan anlaşıldığı üzere verilen herhangi bir çizgenin Erdős-Rényi modeli altındaki bileşen olasılığını hesaplamak için Yeterli istatistikleri kullanmak ve b parametresinin önsel dağılımı olan Beta dağılımının parametrelerini (α, β) sağlamak yeterlidir.

Rastlantısal öbek rastgele çizgeler için marjinal olabilirlik ise şu şekilde hesaplanır:

$$\log p(A) = \log \sum_C \int_B p(A|C, B) p(B) p(C) dC dB \quad (12)$$

Herhangi bir kategori ataması için $p(A|C)$ koşullu olasılığını Erdős-Rényi modelinde olduğu gibi hesaplayabiliriz, çünkü her kategorisi ikilisi $\{r, s\}$ bir öbek oluşturmaktadır ve bu öbek

başlı başına B_{rs} ile ifade edilen bir Erdős-Rényi modelidir:

$$\begin{aligned} \log p(A|C) = & \sum_{r,s} \{ \log \Gamma(\alpha_k + \beta_{rs}) + \log \Gamma(\alpha_{rs} + c_{rs}) \\ & + \log \Gamma(\beta_{rs} + n_{rs} - c_{rs}) - \log \Gamma(\alpha_{rs}) \\ & - \log \Gamma(\beta_{rs}) - \log \Gamma(\alpha_{rs} + \beta_{rs} + n_{rs}) \} \end{aligned} \quad (13)$$

Bu durumda, rastlantısal öbek çizgeleri için marjinal olabilirlik hesaplamak olası tüm kategori atamaları üzerinden toplam almayı gerektirir. Küçük boyutlu çizgeleri için bu hesaplanabilir.

$$p(A) = \sum_C p(A|C)p(C)dC \quad (14)$$

4. Yaklaşık Çıkarım Metodları

Tüm olası kategori atamaların sayısı $O(K^N)$ gibi üssel bir ifade olduğundan, boğum sayısı arttığında kesin olabilirlik hesabı yapmak imkansızlaşır. Bu durumda B ve C parametrelerini kestirerek gerçek marjinal olabilirliğe yaklaşabiliriz. Biz bu çalışmamızda varyasyonel beklenti enbüyütme ve Gibbs örnekleme yöntemlerini uyguladık.

4.1. Varyasyonel Beklenti Enbüyütme Yöntemi

Bu yöntemde model parametrelerinin ardıl olasılığına varyasyonel bir dağılımla yaklaşacağız.

$$q(B, C) \propto p(B, C|A) \quad (15)$$

Yaklaşık dağılımla hesaplayacağımız yaklaşık marjinal olasılığın alt sınırını şu şekilde ifade edebiliriz [3]:

$$\begin{aligned} \log p(A) \geq & \langle \log p(A, B, C) \rangle_{q(B, C)} \\ & + \langle \log q(B, C) \rangle_{q(B, C)} \end{aligned} \quad (16)$$

$q(B, C)$ dağılımını hesaplayabilmek için önce dağılımı faktörize ediyoruz:

$$q(B, C) = \prod_{r,s} q(B_{rs}) \prod_i q(C_i) \quad (17)$$

$q(C_i)$ ayrık bir dağılım olduğundan, her bir $C_{i,r}$ olasılığını ayrı ayrı hesaplayacağız, $q(B_{rs})$ için ise Beta dağılımını seçiyoruz. $q(B, C)$ dağılımını bulmak için gerçek $p(B, C|A)$ dağılımı ile arasındaki Kullback-Leibler mesafesini enküçültüğümüzde aşağıdaki güncelleme denklemlerini elde ediyoruz:

$$q(B_{rs}) \sim \text{Beta}(\alpha_{B_{rs}}, \beta_{B_{rs}}) \quad (18)$$

$$\alpha_{B_{rs}} = \alpha_{rs} + \sum_{i,j} \langle C_{i,r} \rangle \langle C_{j,s} \rangle A_{ij} \quad (19)$$

$$\beta_{B_{rs}} = \alpha_{rs} + \sum_{i,j} \langle C_{i,r} \rangle \langle C_{j,s} \rangle (1 - A_{ij}) \quad (20)$$

$$\begin{aligned} \log q(C_i) = & + \sum_{i,j} \sum_{r,s} C_{i,r} \langle C_{j,s} \rangle (A_{ij} \langle \log B_{rs} \rangle \\ & + (1 - A_{ij}) \langle \log(1 - B_{rs}) \rangle) \\ & + \sum_r C_{i,r} \gamma_r \end{aligned} \quad (21)$$

4.2. Gibbs Örnekleme Yöntemi

Bir Markov zinciri Monte Carlo yöntemi olan Gibbs örnekleme yönteminde [4] ise $p(B, C|A)$ dağılımından bir miktar $\{C, B\}$ eşleri örnekliyoruz. Daha sonra bu eşleri kullanarak Chib yöntemi [5] ile marjinal olabilirlik hesaplayacağız. Gibbs düğümlerinde her bir parametreyi diğer tüm parametrelerin o anki değerlerinin ve gözlemlenen verinin tam koşullu olasılıklarından örneklediğimiz takdirde elde ettiğimiz $p(B, C|A)$ eşleri $p(B, C|A)$ dağılımından örneklenmiş olur. C parametresinin tam koşullu olasılığını Bayes kuralı kullanarak yazacak olursak:

$$C_{i,r} \sim p(C_{i,r}|A, B, C_{-i}) \quad (22)$$

$$\propto p(A|B, C_{-i}, C_{i,r} = 1) \quad (23)$$

$$=^+ \exp \left(\sum_{j=1}^N \sum_{s=1}^K C_{j,s} W_{i,j}^{r,s} \right) \quad (24)$$

Burada C_{-i} ifadesi i boğumu dışında kalan tüm boğumları ifade etmektedir. $B_{-(r,s)}$ de aynı şekilde B_{rs} dışındaki B değerlerini ifade ettiğinde, B_{rs} için tam koşullu olasılığı aşağıdaki gibi ifade edebiliriz:

$$B_{rs} \sim p(B_{rs}|A, B_{-(r,s)}, C) \quad (25)$$

$$\propto p(A|B_{-(r,s)}, B_{rs} = \sigma, C) \quad (26)$$

Buradan B değerlerini parametreleri 27. ve 28. denklemlerde verilen bir Beta dağılımdan örnekleyebileceğimize ulaşabiliriz.

$$\alpha = \sum_{i,j} A_{ij} C_{i,r} C_{j,s} + 1 \quad (27)$$

$$\beta = \sum_{i,j} (1 - A_{ij}) C_{i,r} C_{j,s} + 1 \quad (28)$$

Gibbs yöntemi kullanarak örneklediğimiz G adet eş $\{B^{(g)}, C^{(g)}\}_{g=1}^G$ ile ifade edecek olursak, marjinal olabilirlik hesabını Chib metoduna dayanarak şu şekilde gösterebiliriz:

$$\log p(A) = \log p(A|B^*) + \log p(B^*) - \log p(B^*|A) \quad (29)$$

Bu denklem tüm B değerleri için doğrudur. B^* ise örnekleğimiz tüm B değerleri içerisinde olabilirlik değerine bakarak seçtiğimiz B değerini ifade etmektedir. Denklemdaki terimleri elimizdeki örnekleri kullanarak Monte Carlo yaklaşımıyla şu şekilde hesaplayabiliriz:

$$p(A|B^*) = G^{-1} \sum_{g=1}^G p(A|B^*, C^{(g)}) \quad (30)$$

$$(31)$$

$p(B^*|A)$ için ise B^* değerini sabit tutarak yeni $C^{(g)}$ örnekleme ihtiyacı duyuyoruz.

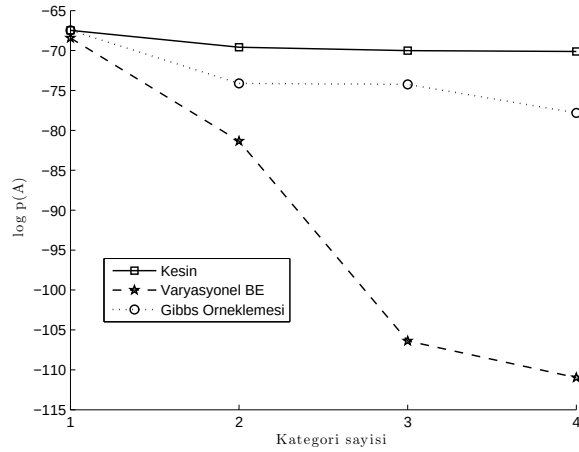
$$p(B^*|A) = G^{-1} \sum_{g=1}^G p(B^*|A, C^{(g)}) \quad (32)$$

$$p(B^*|A, C^{(g)}) = \frac{p(A|B^*, C^{(g)})p(B^*)}{p(A|C^{(g)})} \quad (33)$$

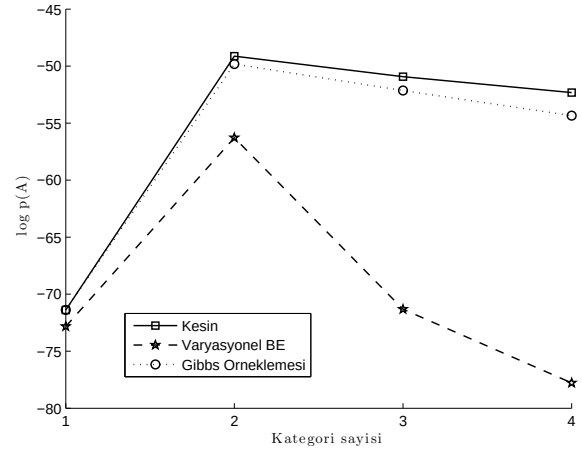
5. Sonuçlar ve Vargılar

Yukarıda 1. ve 2. şekilde görülen çizgeler için kesin marjinal olabilirlik değerlerini hesapladık. Daha sonra varyasyonel metod ve gibbs örnekleme ile bu değerlere yaklaşılmaya çalıştık. 3. ve 4. grafiklerde sırasıyla Erdős-Rényi modelinden üretilen 1. çizge ve rastlantısal öbek modelinden üretilen 2. çizge için marjinal olabilirlik değerleri görülmektedir. Erdős-Rényi modeli ile oluşturulan çizge için alınan sonuçlarda, her 3 yöntemin de en yüksek marjinal olabilirliği 1. kategori değeri için verdiğini, ve diğer kategori değerleri için benzer bir eğimle daha düşük olabilirlik verdiklerini gözlemliyoruz. 2 kategorili rastlantısal öbek modelden oluşturulmuş çizge için ise tüm yöntemler en yüksek olabilirliği 2 kategori için verdiler.

Gibbs örneklemesinin gerçek marjinal olabilirliğe daha yakın sonuçlar çıkarması, varyasyonel yaklaşım yöntemimizi daha iyileştirmemiz doğrultusunda bir ip ucu verdi. Gerçekte çizge boğumlarının kategori dağılımları birbirleriyle sıkı bir korelasyon içerisinde olduğundan, $q(C)$ dağılımını faktörize etmenin bu sonucu doğurduğunu düşünüyoruz. Bu faktörizasyonu $q(C_i, C_j)$ gibi ikili veya daha büyük bloklarla yapmayı ilerisi için düşünmekteyiz. Ayrıca karışık rastantısal öbek modelleri [1] de bu model seçimine dahil etmeyi planlamaktayız.



Şekil 3: Çizge 1 için hesaplanan olabilirlik değerleri grafiği



Şekil 4: Çizge 2 için hesaplanan olabilirlik değerleri grafiği

- [5] Chib S., "Marginal Likelihood From the Gibbs Output". Journal of the American Statistical Association. 1995; 90(432).

6. KAYNAKÇA

- [1] Goldenberg A., Zheng A. X., Fienberg S. E., Airolidi E. M., "A survey of statistical network models", Foundations and Trends in Machine Learning, 2(2):1-117, 2009.
- [2] B. Olding ve P. J. Wolfe. "Inference for graphs and networks: Extending classical tools to modern data." Submitted for publication, 2009.
- [3] Bishop C., "Pattern Recognition and Machine Learning (Information Science and Statistics)", Springer, 2007.
- [4] Liu, J. S., "Monte Carlo Strategies in Scientific Computing", Springer, 2001.