RASSAL ÇİZGE SERİLERİNDE ÇOKLU DEĞİŞİM NOKTASI ANALİZİ
MULTIPLE CHANGE POINT ANALYSIS IN RANDOM GRAPH SERIES

Türkan Hamzaoğlu[1]*, Barış Kurt*[1]*, A. Taylan Cemgil*[1]

[1]Bilgisayar Mühendisliği Bölümü
Boğaziçi Üniversitesi, İstanbul
turkan.hamzaoglu@boun.edu.tr, baris.kurt@boun.edu.tr, taylan.cemgil@boun.edu.tr

ÖZETÇE

Çizgeler süreçlerin modellenmesinde kullanılan önemli matematiksel araçlardır. Bu alanda önemli bir problem, zamanla üretici süreç parametrelerinde meydana gelen değişimleri çıkarımlamaktır. Bu çalışmada, rassal öbek çizgeleri serilerinde çoklu değişim noktalarını çıkarımlama problemi ele alındı. Zaman serilerinde çıkarım yapmak için en sık kullanılan algoritmalardan biri ileri-geri algoritmasıdır. Bu algoritmanın çizge değerli modellerde hesaplama karmaşıklığını azaltmak için, geri yönlü mesaj kısmı değiştirilip geriye doğru Monte Carlo örneklemesi uyarlandı. Yapılan testlerle tasarlanan algoritmanın gerçeğe uygun çıkarımlar yaptığı gözlemlendi.

ABSTRACT

Graphs are important mathematical tools for modelling processes. An important issue in this area is to infer the changes that occur in the underlying generative process. In this work, inference of multiple change points in stochastic block graph time series is studied. A well-known algorithm for inference in time series is the forward-backward algorithm. In order to decrease computational complexity of this algorithm in graphical models, backward smoothing part is replaced with backward Monte Carlo sampling. With the experiments, it is observed that modified algorithm gives result in accordance with the real data.

1. GİRİŞ

Rassal çizgeler, birçok alanda process ve veri modellemelerinde ve veri analizinde kullanılmaktadır. En yaygın kullanım alanlarının başında telekominikasyon ağlarının incelenmesi, ulaşım şebekesinin modellenmesi, mobil telefon ağlarının incelenmesi, proteinler arası etkileşimlerin modellenmesi, besin zinciri gibi biyolojik ağların analizi; facebook, twitter sosyal ağların analizi; netflix, amazon gibi kullanıcı tercihlerinin incelenmesidir. Bu örnekler daha da çoğaltılabilir.

Bu çalışma Boğaziçi Üniversitesi BAP Koordinatörlüğü 5723 numaralı proje tarafından desteklenmektedir.

978-1-4673-0056-8/12/$26.00 ©2012 IEEE

Yukarıdaki ve benzeri analizler için değişik çizge modelleri kullanılmaktadır. Literatürde en çok yer alan modeller: Erdös-Reyni-Gilbert modeli, sosyal ağlar için P_1 ve P_2 modelleri, üssel rassal çizge modelleri, rassal blok modelleri, tercihli bağlanma modeli, çoğaltmalı bağlanma modeli ve türevleri olarak sayılabilir[1].

Bu makalede, Rassal Blok Model Çizgeleri ve bunların dinamik yani zaman içerisinde değişimi üzerinde çalışıldı. Bu modelde düğümlerin örtülü kategorilere dahil oldukları ve düğümler arasındaki ilişkilerin, bu örtülü kategorilere göre oluştuğu varsayılır[2]. Örnek olarak, bir basit kullanıcı-servis çizgesi düşünelim. Kullanıcıların tercih ettikleri servisler kullanıcı ve servis kategorilerine göre belirlenmektedir. Modelle ilgili matematiksel ayrıntılar ileriki bölümlerde verilecektir.

Günümüzde durağan veri analizi kadar, dinamik veri analizi de önem kazandı. Örnek olarak, ilgili süreçlerdeki değişimi takip etmek ve değişime göre önlem almak ya da eylemde bulunmak ilgili kurumlar için önemlidir. Rassal blok modellerinden örnek vermek gerekirse, zaman içinde müşteri ve servis tercihleri arasındaki değişimi farketmek, servis sağlayıcıları için önemli olacaktır. Bu ve benzeri zaman içindeki değişim noktalarını çıkarımlamak için rassal çizge modelleri zaman serisi modelleri olarak tasarlandı.

2. RASSAL ÖBEK ÇİZGELERİ ve ÇOKLU DEĞİŞİM NOKTALARI

Bu bölümde önce rassal öbek çizgelerinin matematiksel modeli incelenecek, sonrasında da rassal öbek çizgileri zaman serilerinde çoklu değişim noktalarının nasıl modellendiğine değineceğiz. Çoklu değişim noktalarının çıkarımlanmasında ise saklı Markov modellerinin analizinde kullanılan *ileri yönde filtreleme ve geri yönde düzeltme* (forward filtering-backward smoothing) algoritmasının Monte Carlo örneklemesine uyarlanması olan *ileri yönde filtreleme ve geri yönde örnekleme* (forward filtering-backward sampling) algoritması kullanılmıştır.

2.1. Rassal Öbek Çizgeleri

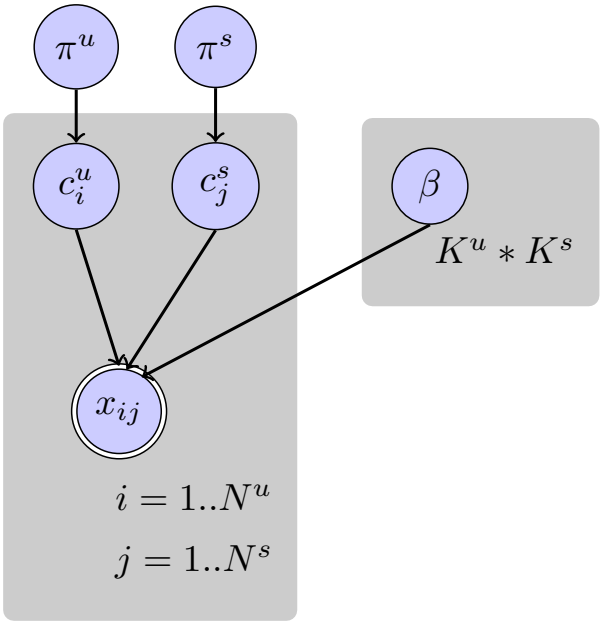
Temel rassal öbek çizgelerinde herbir düğüm örtülü alt bir kategoriye aittir. Düğümler arası bağlantılar da bu kategorilere göre belirlenir. Literatürde, bu modelin çeşitlemeleri tanımlanmıştır. Bazı modellerde, düğümler arasındaki kenarlar alt kategorilere ek olarak, düğümlere ya da düğümlerin çeşitli özelliklerine de bağlanabilir. Bu makalede, temel rassal öbek çizgelerini ele alacağız ama yukarıda belirtilen çeşitlemeler de kolayca uyarlanabilir. Şekil 1'de modelin Bayesçi çizgelerle betimlenmesi gösterilmektedir. Önceki kullanıcı-servis örneği baz alınarak, altta yatan matematiksel model şu şekilde betimlenebilir:

$$
\begin{aligned}
&N^u = \textit{Kullanıcı sayısı}\\
&K^u = \textit{Kullanıcı kategori sayısı}\\
&N^s = \textit{Servis sayısı}\\
&K^s = \textit{Servis kategori sayısı}\\
&k \in \{1,\ldots,K^u\} = \textit{Kullanıcı kategori endeksi}\\
&l \in \{1,\ldots,K^s\} = \textit{Servis kategori endeksi}\\
&p \in \{1,\ldots,N^u\} = \textit{Kullanıcı endeksi}\\
&q \in \{1,\ldots,N^s\} = \textit{Servis endeksi}\\
&C^u_p \sim Multn(\pi^u) = \textit{p kullanıcısının kategorisi}\\
&C^s_q \sim Multn(\pi^s) = \textit{q servisinin kategorisi}\\
&B_{kl} \sim Beta(\alpha,\beta) = \textit{k kategorili kullanıcı ile}\\
&\qquad \textit{l kategorili servis arasında ilişki olasılığı}\\
&Y_{pq} = \begin{cases} 1 & \textit{eğer p kullanıcısı q servisini kullanıyorsa}\\ 0 & \textit{diğer şartlarda}\end{cases}\\
&\qquad \sim Bern(C^u_p B (C^s_q)^T) = \textit{kullanıcılar ve servisler}\\
&\qquad \textit{arasındaki komşuluk matrisi}
\end{aligned}
\tag{1}
$$

Tüm modelin bileşik olasılık dağılımı şu şekilde elde edilir.

$$
\begin{aligned}
p(Y,B,&C^u_{1:N^u},C^s_{1:N^s}) = p(C^u_{1:N^u}) * p(C^s_{1:N^s})\\
&* p(B) * p(Y|\beta, C^u_{1:N^u}, C^s_{1:N^s})\\
&= \prod_p^{N^u}\prod_k^{K^u}(\pi^u_k)^{C^u_{pk}} * \prod_q^{N^s}\prod_l^{K^s}(\pi^s_l)^{C^s_{ql}} *\\
&\prod_k^{K^u}\prod_l^{K^s}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}B_{kl}^{(\alpha-1)}(1-B_{kl})^{(\beta-1)}\\
&* \prod_p^{N^u}\prod_q^{N^s}\prod_k^{K^u}\prod_l^{K^s}(B_{kl}^{Y_{pq}}(1-B_{kl})^{(1-Y_{pq})})^{C^u_{pk}*C^s_{ql}}
\end{aligned}
\tag{2}
$$

Bu model için düğümlere kategori atama (C_u ve C_s) ve bağlantı parametrelerini (B) çıkarım yöntemleri geçen yılki makalemizde incelenmişti [3]. Burada dikkat edilmesi gereken önemli nokta, düğümlerin kategorileri bilindiğinde, rassal öbek modeli koşullu bağımsız olarak



Şekil 1: Rassal Öbek Çizge Modeli

$C_u * C_s$ tane alt Erdos-Reyni çizgesi olarak ifade edilebilmesidir.

Herhangi bir A Erdos-Reyni matrisini alalım. Bu matriste kenarlar arası bağlantı olasılığı $Bernoulli(b)$ olasılık dağılımı ile belirlenmiş olsun. b parametresinin eşlenik önsel dağılımı da $Beta(\alpha,\beta)$ olsun. Bu durumda A matrisinin marjinal olasılığı aşaıdaki şekilde kolayca bulunabilir:

$$p(A) = \int_b p(b)p(A|b)db \tag{3}$$

$$c = \sum_{i,j}^{N} A_{ij} \quad \text{(toplam bağ sayısı)} \tag{4}$$

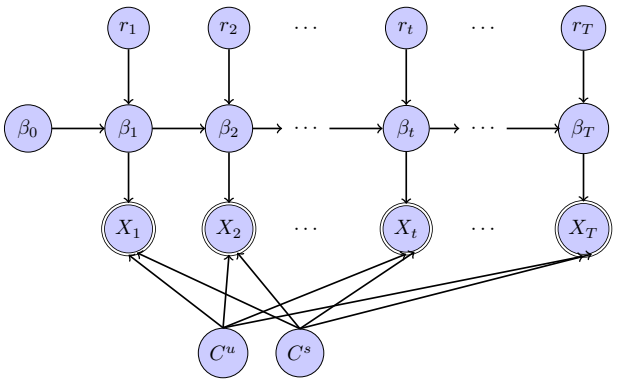
$$n = \sum_{i,j}^{N} 1 \quad \text{(olabilecek tüm bağların sayısı)} \tag{5}$$

$$p(b) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}b^{\alpha-1}(1-b)^{\beta-1} \tag{6}$$

$$\log p(A) = \log \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}b^{\alpha+c-1}(1-b)^{\beta+n-c-1}db \tag{7}$$

$$
\begin{aligned}
&= \log\Gamma(\alpha+\beta) + \log\Gamma(\alpha+c)\\
&\quad + \log\Gamma(\beta+n-c) - \log\Gamma(\alpha)\\
&\quad - \log\Gamma(\beta) - \log\Gamma(\alpha+\beta+n)
\end{aligned}
\tag{8}
$$

Yukarıdaki denklemlere göre, A matrisinin bağlantı parametresi b'nin çıkarımı oldukça basittir.



Şekil 2: Çoklu Değişim Noktaları Modeli

2.2. Rassal Öbek Çizgelerinin Zaman Serilerinde Çoklu Değişim Noktalarının Çıkarımı

Çoklu değişim noktaları modelinde, T uzunluğundaki zaman serisi için değişim noktalarının sayısını ve yerleri bilinmiyordur ve amaç bu değişim noktalarının yerlerini çıkarımlamaktır. Bu modelde, herhangi bir t anında değişim olduysa, rassal öbek çizge modelimizdeki bağlantı parametrelerinin önsel dağılımından tekrar çekildiği, eğer değişim olmadıysa $t-1$ bağlantı parametreleriyle aynı olduğu varsayılır. Varsayılan diğer önemli bir nokta da, öbek çizgelerindeki düğümlerin kategorilerinin bilinmiyor ve T süresince değişmiyor olmasıdır. Her t anında öbek çizgeleri ilgili bağlantı parametlerine ve altta yatan kategori eşlemesine göre yeniden oluşturuluyor. Şekil 2'de çoklu değişim noktaları modelimizin Bayesçi çizgelerle temsilini görüyoruz.

Matematiksel olarak üretici modeli şu şekilde tanımlayabiliriz:

$$r_t \sim Be(p) \tag{9}$$

$$\beta_t \sim [r_t = 0]\delta(B_t = B_{t-1}) \tag{10}$$

$$B_t \sim [r_t = 0]\delta(B_t = B_{t-1}) + [r_t = 1]\prod_k^{K^u}\prod_l^{K^s} Beta(\alpha, b) \tag{11}$$

$$C_p^u \sim Multn(\pi^u) \tag{12}$$

$$C_q^s \sim Multn(\pi^s) \tag{13}$$

Herhangi bir t anındaki komşuluk matrisinin (Y_t) koşullu olasılığı önceki bölümde verildiği gibi hesaplanır:

$$\begin{aligned} Y_t|B_t, C^u, C^s = &\prod_p^{N^u}\prod_q^{N^s}\prod_k^{K^u}\prod_l^{K^s}(B_{t,kl}^{Y_{t,pq}} \\ &(1 - B_{t,kl})^{(1-Y_{t,pq})})^{C_{pk}^u * C_{ql}^s} \end{aligned} \tag{14}$$

2.2.1. İleri-Geri Algoritması

İleri-geri algoritması zaman serileri ve saklı Markov modellerinde anlık parametre çıkarımlarında çok sık kullanılan bir tam çıkarım yöntemidir[4]. Bizim modelimizde de $p(r_t, B_t|Y_{1:T})$ koşullu dağılımları bu algoritma ile hesaplanabilir. Algoritmanın adımlarını kısaca tanımlayalım.

Alfa (ileri yönlü) mesajları:

$$\begin{aligned} \alpha_{0|0} &= p(B_0) \\ t &= 1..T \\ \alpha_{t|t-1} &= p(r_t, B_t, Y_{1:t-1}) \\ \alpha_{t|t} &= p(r_t, B_t, Y_{1:t}) \end{aligned} \tag{15}$$

Beta (geri yönlü) mesajları:

$$\begin{aligned} \beta_{T|T} &= p(Y_T|r_T, B_t) \\ t &= T-1, ..1 \\ \beta_{t|t+1} &= p(Y_{t+1:T}|r_t, B_t) \\ \beta_{t|t} &= p(Y_{t:T}|r_t, B_t) \end{aligned} \tag{16}$$

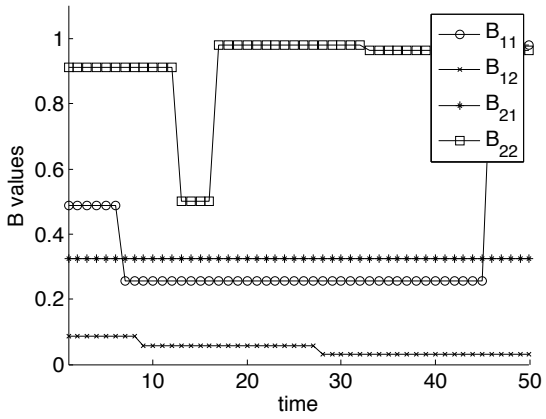
Değişim noktalarının sonsal dağılımları da alfa ve beta mesajları kullanılarak şu şekilde elde edilir:

$$\begin{aligned} p(r_t, B_t|Y1:T) \sim &p(Y_{1:T}, r_T, B_t) \\ = &p(Y_{1:t-1}, r_t, B_t)p(Y_{t:T}|r_t, B_t, Y_{1:t-1}) \\ = &p(Y_{1:t-1}, r_t, B_t)p(Y_{t:T}|r_t, B_t) \\ = &\alpha_{t|t-1}\beta_{t|t} \end{aligned} \tag{17}$$

2.2.2. İleri Yönlü Filtreleme - Geri Yönlü Örnekleme Algoritması

İleri-geri algoritması çıkarımlanmak istenen parametrelerin tüm olası değerleri için her bir zaman adımında hesaplamaları gerektirdiğinden, hesaplama karmaşıklığı olarak üsseldir. Hesaplama karmaşıklığını azaltmak için alfa-beta algoritmasının geri yönlü mesajlarının tümünü hesaplamak yerine, T anındaki sonsal dağılımdan, T zamanı için değişim noktası örneklenir (r_T). Örnek değeri için beta mesajını $(\beta_{t-1|t-1})$ hesaplayıp $t-1$ anı için sonsal dağılımları çıkarımlanır. Örneklemeye bu şekilde sonsal dağılımlardan devam edilir. Bu ufak değişiklikle hesaplama karmaşıklığı geri yönde doğrusal olur.

Algoritmanın detayları aşağıdaki sözde kodla açıklanabilir:



Şekil 3: Zaman dilimlerine göre bağlantı parametrelerinin değişimi

Algoritma 1: İLERI FILTRELEME-GERI ÖRNEKLEME(.)

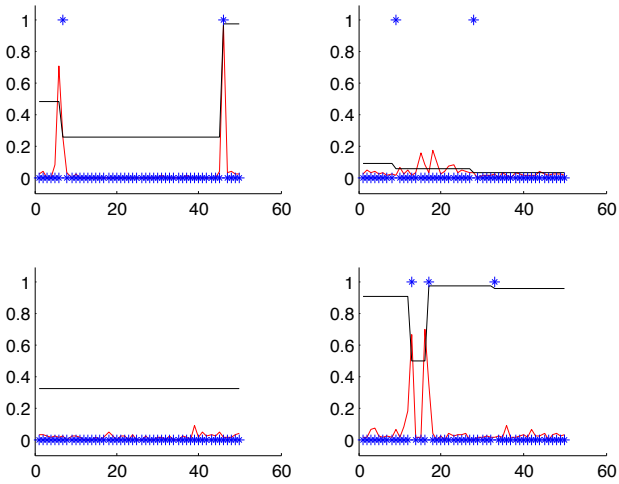
for each $p \in \{1, .., N^u\}$
  do $C_p^u \sim Multn(\pi^u)$ örnekle
for each $q \in \{1, .., N^s\}$
  do $C_q^s \sim Multn(\pi^s)$ örnekle
for each $t \in \{1, .., T\}$
  do $r_t \sim Be(p)$ örnekle

Gibbs Örneklemesi:

for each $i \in \{1, .., epoch\}$
  do
    for each $t \in \{1, .., T\}$
      do
        for each $k \in \{1, .., K^u\}$
          do for each $l \in \{1, .., K^s\}$
          do $A_{t,kl}^i$ = k ve l kategorilerine dahil düğümlerden oluşan alt çizge
        İleri Süzgeçleme :
        for each $A_{t,kl}^i$
          do $\alpha_{t,kl}^i(r_{t,kl})$ mesajlarını hesapla
        Geri Örnekleme:
        for each $t \in \{T, T-1.., 1\}$
          do
            s_t :düzeltilmiş sonsal olasılık.
            $r_t \sim s_t$ t dilimi değişim nokt.
            $\beta_t(r_t)$ beta potansiyelleri
for each $p \in \{1, .., N^u\}$
  do $C_p^u \sim p(C_p^u | Y_{1:T}, C_{p^-}^u, C^s)$ örnekle
for each $q \in \{1, .., N^s\}$
  do $C_p^u \sim p(C_q^s | Y_{1:T}, C_{q^-}^s, C^u)$ örnekle

3. SONUÇLAR

Önerilen yeni algoritmayı test etmek için daha önce belirtilen zaman serisi modeline göre 50 zaman dilimi için veri üretildi. Üretici modelde 2 kullanıcı kategorisi ve 2 servis kategorisi olduğu varsayıldı. Her bir zaman diliminde değişim noktası olma olasılığı 0.04 olarak alındı. Her bir düğümün herhangi bir kategoriye ait olma olasılığı 0.5 olarak kabul edildi. Verideki gerçek değişim noktalarını, düğümlerin kategorilerini ve kategoriler arası bağlantı parametreleri şekil 3'te verilmiştir.



Şekil 4: Örneklenen değişim noktalarının sonsal dağılımları. Düz çizgiler bağlantı parametrelerinin değişimini, mavi yıldızlar gerçek değişim noktalarını, kırmızı grafikler de çıkarımlanan sonsal dağılımları gösteriyor.

Önerilen algoritmanın ürettiği sonuçlar da şekil 4'de görülmektedir. Bu şekilden de anlaşıldığı gibi, Gibbs' örneklemesi içinde geri yönlü örneklemelerden elde edilen noktalar, değişim noktaları için sonsal dağılım oluşturmaktadır. Gerçek değişim noktaları da, bu sonsan dağılımın en yüksek olasılık değerleriyle örtüşmektedir. Böylelikle önerilen algoritmanın daha az işlem karmaşıklığıyla tutarlı çıkarımlar yaptığı gözlemlendi.

4. KAYNAKÇA

[1] Goldenberg, A. and Zheng, A. X. and Fienberg, S. E. and Airoldi, E. M.,"A survey of statistical network models",arXiv,2009,6481222.

[2] Airoldi, Edoardo M. and Blei, David M. and Fienberg, Stephen E. and Xing, Eric P., "Combining stochastic block models and mixed membership for statistical network analysis", ICML'06, 57–74, 2007.

[3] Kurt, Barış and Cemgil, A. T., "Rastlantısal Öbek Çizgeler İçin Bayesçi Model Seçimi", SIU 2011.

[4] Bishop C., "Pattern Recognition and Machine Learning (Information Science and Statistics)", Springer, 2007.