Olasılıksal Uzman Sistemlerde Soru Sıralama Stratejileri
Query Ranking Strategies in Probabilistic Expert Systems

Hıdır Yüzügüzel*, Ali Taylan Cemgil†, Emin Anarım*
Elektrik ve Elektronik Mühendisliği Bölümü*
Bilgisayar Mühdendisliği Bölümü†
Boğaziçi Üniversitesi
{hidir.yuzuguzel, taylan.cemgil, anarim}@boun.edu.tr

***Özetçe* —Bir çok alanda özniteliklerin sayısı oldukça yüksektir. Örneğin tıp alanında kullanılan olasılıksal uzman sistemlerde semptomların sayısı 1000'ler mertebesindedir. Burada tıbbi tanıya ulaşmak için bütün semptomları sorgulamak pratik olmadığından sıralama seçimi önem kazanmaktadır. Bu çalışmada, olasılıksal uzman sistemlerde 3 tane soru sıralama stratejisi önerilmekte ve bu stratejilerin yapay veriler üzerindeki başarımları değerlendirilmektedir.**

Anahtar Kelimeler—tıbbi tanı, sıralı tanı , bağıl-entropi

***Abstract*—The number of features are quite high in many fields. For instance, the number of symptoms are around thousands in probabilistic medical expert systems. Since it is not practical to query all the symptoms to reach the diagnosis, query choice becomes important. In this work, 3 query ranking strategies in probabilistic expert systems are proposed and their performances on synthetic data are evaluated.**

Keywords—medical diagnosis, sequential diagnosis, relative-entropy

I. GİRİŞ

Tıbbi tanı (Medical diagnosis) bir örüntü sınıflandırma (pattern classification) problemi olarak düşünülebilir. Temel olarak örüntü sınıflandırma, verilen bir nesneyi bilinen k sınıftan herhangi birine atamayı ele alır. Sıralı sınıflandırma (online classification) ise sıralı bir biçimde ilerler [1]. Sıralı sınıflandırmada [2], öznitelikler teker teker sınanır, sonsal olasılıklar hesaplanır ve sınamanın devam edeceğinin ya da son bulacağının kararı verilir. Eğer sınama devam ederse, sınama için bir sonraki öznitelik seçilir. Aksi takdirde, sınıflandırma yapılır. Örneğin tıp alanında kullanılan olasılıksal uzman sistemlerde sınıflandırma tıbbi tanıya, öznitelikler ise semptomlara karşılık gelmektedir.

Bir çok alanda özniteliklerin sayısı yüksektir. Örneğin tıp alanında yaygın bir şekilde kullanılan QMR-DT [3] adlı uzman sistemde yaklaşık olarak 600 hastalık ve 4000 semptom bulunmaktadır. Burada çıkarım için bütün semptomları sorgulamak pratik değildir. Bu yüzden sıralama seçimi önem kazanmaktadır.

978-1-4673-5563-6/13/$31.00

Bir doktorun belirli semptomların varlığı hakkında sorular sorduğu ya da tıbbi sınamaları tavsiye ettiği tipik bir tanı sürecini düşünelim. Doğal olarak bir doktor, hastasının semptomları hakkında ne kadar fazla bilgi alabilirse tanı daha kesin ve muhtemelen daha doğru olacaktır. Doğruluk için, doktor olası tüm testleri tavsiye edebilir ancak bu stratejinin maliyeti yüksek olacaktır. Burada söz konusu olan maliyet hastanın rahatsızlığı, zaman, para ya da bunların bir kombinasyonu cinsinden ölçülebilir. Arzu edilen ise doğru tanıya mümkün olan en az soruyu sorarak ulaşabilmektir. Bu da, en bilgi verici soruları sormakla mümkündür.

En bilgi verici soruyu seçme problemi başta tıbbi tanı, karar analizi ve öznitelik seçme olmak üzere birçok yapay öğrenme probleminde karşımıza çıkmaktadır. [4]'te en bilgi verici sorular altkümesini seçmek için koşullu entropideki düşüş bir yöntem olarak önerilmiştir. Önerilen algoritma döngüsel inanç yayılımına (loopy belief propagation) dayalı olup, o ana kadar sorulan sorulara verilen yanıtları hesaba katıp karşılıklı bilgi miktarındaki kazancı hesaplayarak soruları sıralı bir biçimde seçmektedir. [5]' te sıralama tabanlı aç gözlü bir algoritma önerilmiştir. Bu algoritma, sıralama tabanlı çıktının ROC eğrisinin altında kalan alanı enbüyüten soruları sıralı bir biçimde seçmektedir.

Bir soru sıralama stratejisi tam tanının doğruluğuna hızlı bir şekilde ürettiği birkaç iyi seçilmiş soruyla ulaşabilirse etkilidir. Bu çalışmada üç tane strateji incelenmiştir: bilgi-kuram tekniği, bağıl entropi (relative-entropy) tabanlı stratejisi olarak da bilinir, semptomlara dayalı bir strateji ve hastalıklara dayalı bir strateji. Bu üç strateji de semptomlar listesinden rastgele soru seçip soran bir strateji ile karşılaştırılmıştır.

Bu çalışmada, Bölüm 2'de soru sorma stratejileri ayrıntılı bir biçimde açıklanmıştır. Deneysel çalışmalar Bölüm 3'te, vargılar ise Bölüm 4'te verilmektedir.

II. SORU SIRALAMA STRATEJİLERİ

A. Bilgi-kuram (Bağıl-Entropi tabanlı) Stratejisi

Bağıl-entropi tabanlı strateji [6] soruları entropiyi azaltmadaki etkinliğine göre seçmektedir. Bir başka deyişle bu strateji tanının Shannon entropisindeki düşüşü enbüyüten soruları seçmektedir. Bağıl-entropi, $s = (s_1, \ldots, s_M)$ semptom

vektörü olmak üzere $s(\sigma(n))$'in bir tanı hakkında sağladığı ek bilginin ölçüsüdür. Bu bağlamda, $n = 1 \dots M$ (toplam semptom sayısı) olmak üzere $\sigma(n)$ n'inci semptomun sorulma sırasının indisini belirten permütasyondur. p_1 ve p_2 birer olasılık dağılımı olarak tanımlanmıştır:

$$p_1 = p(d_i|s(\underbrace{\sigma(1), \dots, \sigma(n-1)}_{\sigma'}, \boldsymbol{\sigma(n)})) = p(d_i|s(\sigma', \sigma(n))) \tag{1}$$

$$p_2 = p(d_i|s(\sigma(1), \dots, \sigma(n-1)) = p(d_i|s(\sigma')) \tag{2}$$

p_2, d_i hastalığının $s(\sigma(n))$ semptomu gözlemlenmeden önceki olasılığı, yani d_i'nin önsel dağılımı, ve p_1, d_i hastalığının $s(\sigma(n))$ semptomu gözlemlenmdikten sonraki olasılığı, yani d_i'nin marjinal sonsal olasılığıdır. KL ıraksaklığı önsel dağılımdan sonsal dağılıma hareket ederkenki bilgi kazancının ölçüsü olarak kullanılmıştır. p_1 ve p_2 arasındaki KL ıraksaklığı şu şekilde tanımlanmıştır:

$$\mathcal{D}_{KL}(p_1||p_2) = \sum_i p(d_i|s(\sigma', \sigma(n))) \ln \left(\frac{p(d_i|s(\sigma', \sigma(n)))}{p(d_i|s(\sigma'))} \right) \tag{3}$$

$$= H(s(\sigma(n))) \tag{4}$$

Sorulacak en iyi soru $s(\sigma(n))$ enbüyük beklenen bağıl-entropiyi verendir ve beklenen bağıl-entropi şu şekilde hesaplanmıştır:

$$\mathbb{E}[H(s(\sigma(n)))] = \sum_{s(\sigma(n)) \in \{var, yok\}} p(s(\sigma', \sigma(n)))H(s(\sigma(n))) \tag{5}$$

$p(s(\sigma', \sigma(n)))$ olasılığın açılım kuralından hesaplanabilir:

$$p(s(\sigma', \sigma(n))) = \sum_d p(s(\sigma', \sigma(n))|d)p(d) \tag{6}$$

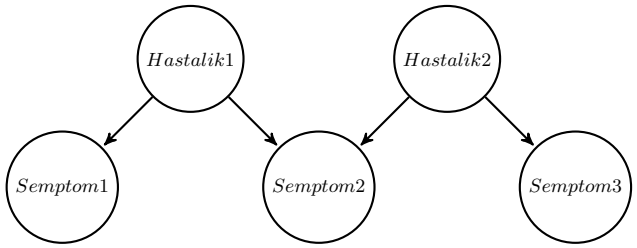
$p(d)$, $d = (d_1, \dots, d_N)$ hastalık vektörünün olasılığıdır ve şu şekilde hesaplanmıştır:

$$p(d) = \prod_i (1 - \pi_i)^{\mathbf{1}\{d_i = 0\}} \pi_i^{\mathbf{1}\{d_i = 1\}} \tag{7}$$

Denklem (7)'deki $\mathbf{1}\{.\}$ gösterge fonksiyonudur ve içerisindeki terim doğru olduğunda 1'e eşittir.

Basit örnek Şekil 1'deki Bayesçi ağı düşünelim. Model parametreleri:

- Hastalıkların önsel dağılımı eşit ve $\pi = 0.01$. Bu parametre bize hastalıkların ender görüldüğünü söylemektedir.
- Bir semptomun ona yol açan hiçbir hastalık yokken görülmeme olasılığı $\theta_0 = 0.99$. Bu parametre bize bir semptomun ona yol açan hiçbir hastalığı olmadığı halde modellenmeyen (hesaba katılmayan) arka plandaki bir hastalıktan ötürü ender görüldüğünü söylemektedir.
- Bir semptomun ona yol açan hastalığı varken görülmeme olasılığı $\theta = 0.02$.



Şekil 1: 2 hastalık ve 3 semptomlu bir Bayesçi ağ. d_i'den s_j'ye olan bağ olası bir etkiyi gösterir ki bu da $D(j,i)$ etki matrisinde ona karışılık gelen elemanın 1 olmasıdır.

İlk olarak (hiçbir semptom hakkında bilgimiz yokken), bağıl-entropi tabanlı strateji *Semptom2*'yi sorar (Tablo I). Şekil 1'deki Bayesçi ağ simetrik bir ağ olduğundan, *Semptom1* ve *Semptom3* aynı beklenen bağıl-entropiye sahiptirler.

Sorular	Beklenen bağıl-entropi
Semptom1	0.0375241
Semptom2	0.0672948
Semptom3	0.0375241

Tablo I: Henüz gözlem yokken

Semptom2'yi gözlemlediğimizi varsayıp devam ettiğimizde, bağıl-entropi stratejisi *Semptom3*'ü sorar (Tablo II). (Alternatif olarak *Semptom1*'i de sorabilirdi)

Sorular	Beklenen bağıl-entropi
Semptom1	0.0128183
Semptom3	0.0128183

Tablo II: Birinci sorudan sonra

Semptom3'ü gözlemlediğimizi varsayıp devam ettiğimizde, bağıl-entropi stratejisi *Semptom1*'i sorar (Tablo III)

Sorular	Beklenen bağıl-entropi
Semptom1	0.000672796

Tablo III: İkinci sorudan sonra

B. Semptom Tabanlı Bir Strateji

M semptom sayısı ve N hastalık sayısı olmak üzere $D(j,i)$ gibi bir hastalık/semptom matrisi verildiğinde:

- İlk olarak, her bir semptom s_j için o semptoma yol açan hastalıkların sayısı sayılarak $S(s_j)$ puan fonksiyonu hesaplanmaktadır:

$$S(s_j) = \sum_i D(j,i) \tag{8}$$

- Daha sonra, semptomlar puanlarına göre büyükten küçüğe sıralanmaktadır ve (kesin) bir tane σ permütasyonu elde edilmektedir.
- Son olarak, semptomlar σ'ya göre sırasıyla sorulmaktadır.

C. Hastalık Tabanlı Bir Strateji

M semptom sayısı ve N hastalık sayısı olmak üzere $D(j,i)$ gibi bir hastalık/semptom matrisi verildiğinde:

- İlk olarak, her bir hastalık d_i için o hastalığın sebep olduğu semptomların sayısı sayılarak $S(d_i)$ puan fonksiyonu hesaplanmaktadır:

$$S(d_i) = \sum_j D(j,i) \tag{9}$$

- Daha sonra, hastalıklar puanlarına göre büyükten küçüğe sıralanmaktadır.
- İlk olarak, en yüksek puana sahip hastalık seçilmektedir ve o hastalığın yol açtığı semptomlardan daha önceden sorulmamış olanları sırası önemsenmeden sorulmaktadır.
- En yüksek puana sahip hastalığın yol açtığı bütün semptomları sorulduğunda, ikinci en yüksek puana sahip hastalığın yol açtığı semptomlar sorulmaktadır, vs.

III. Deneysel Çalışmalar

Deneysel çalışmalar Linux işletim sistemi üzerinde C++ programlama dili kullanılarak gerçellenmiştir. Deneyler için biri küçük ağlarda diğeri geniş ağlarda kullanılmak üzere iki tane ağ verisi üretildi. Her bir ağ verisi rastgele oluşturulmuş ve %30 sıklık oranına sahip 100 tane ağ yapısını içermektedir. Küçük ağ verisindeki her bir ağda 10 hastalık ve 20 semptom bulunmaktadır. Geniş ağ verisindeki her bir ağda ise 100 hastalık ve 400 semptom bulunmaktadır. Her bir ağ yapısı $j = 1 \dots 100$ olmak üzere N_j ile ifade edilmektedir. Geniş ağ verisi gerçek uygulamalarda karşılaşılabilecek ağları yansıtır. Her iki ağ verisi için de kabul edilen model parametreleri şu şekildedir:

- Bir hastalığın önsel olasılığı $\pi = 0.01$
- Bir semptomun arka plandaki hastalığının olasılığı $1-\theta_0$. Deneylerde $\theta_0 = 0.95$ kabul edildi. Seçilen θ_0 değeri, semptomların sebepsiz yere ya da bilinmeyen, modellenmeyen bir hastalık yüzünden ender görülmesi varsayımına uygundur.
- Kayıp olma olasılığı $\theta = 0.02$: Bir semptomun, ona yol açan hastalığının görülmesine rağmen, kendisinin görülmeme olasılığı.

Deney şu şekilde tasarlandı:

- **Veri üretme** Ya elle ya da önselden rastgele örneklemeyle bir tane d hastalık vektörü sabitleştirildi. Daha sonra her bir semptomdan ileri örneklemeyle, her bir N_j ağı için d'ye göre bir tane s_j semptom vektörü üretildi.
- **Çıkarım** Üretilen s_j semptom vektörünü modele koşullandırarak, her bir N_j ağı için d_j^* gibi son tanı çıkarımı yapıldı. Matematiksel olarak şu şekilde yazılabilir:

$$d_j^* = \arg\max_d p(d|s_j) \tag{10}$$

Bu problem en büyük sonsal (MAP) kestirim problemi ve d_j^* de bu problemin MAP çözümü olarak bilinmektedir.

- **Soru Sıralama Stratejilerinin Karşılaştırılması** Her bir N_j ağı için, uzman sistem n'inci semptomu sorgulayan bir soru seçer. Bu da, $n = 1 \dots M$ (toplam semptom sayısı) ve $\sigma(n)$ n'inci semptomun sorulma sırasının indisini belirten permütasyon olmak üzere, $s_j(\sigma(n))$'e karşılık gelir. $\sigma(n)$ indisi her bir strateji için ayrı değerlendirilmiştir. O ana kadar gözlemlenen semptomlar için, ara tanı $d_j^+(n)$ hesaplandı:

$$d_j^+(n) = \arg\max_d p(d|s_j(\sigma(1), \sigma(2), \dots, \sigma(n))) \tag{11}$$

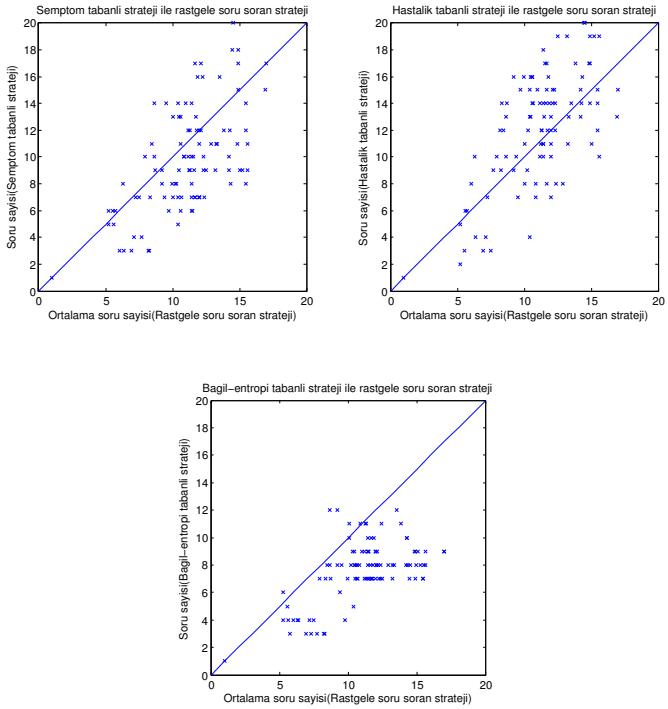
d_j^+ ile d_j^* eşit olduğu anda N_j ağı için o ana kadar sorulmuş olan soru sayısı herhangi bir strateji için t_j ve rastgele soru seçip soran bir strateji için $\underline{t_j}$ olmak üzere kaydedildi. Bu işlem bütün N_j ağları için tekrarlandı. Rastgele soru seçip soran stratejiyi değerlendirmek için, s_j'in 1000 tane rastgele permütasyonu alındı ve $\underline{t_j}$ 1000 tane rastgele permütasyonun ortalamasıyla hesaplandı. Bütün bu işlemler 100 tane rastgele ağ üzerinde yapıldı ve t_j-$\underline{t_j}$ grafiği üretildi. Her bir N_j için hesaplanan t_j ve $\underline{t_j}$ değerleri 2 boyutlu bir uzayda bir noktaya karşılık gelmektedir. Bu noktaların çoğunluğu $x = y$ doğrusunun altında kalırsa, ortalama olarak soru sorma stratejilerinin rastgele soru seçip soran stratejiden ortalama soru sorma sayısı açısından daha üstün olduğunu göstermektedir.

Küçük ve geniş ağ verisi üzerinde yapılan deneyin sonuçları verilmektedir (Şekil 2, Şekil 3).

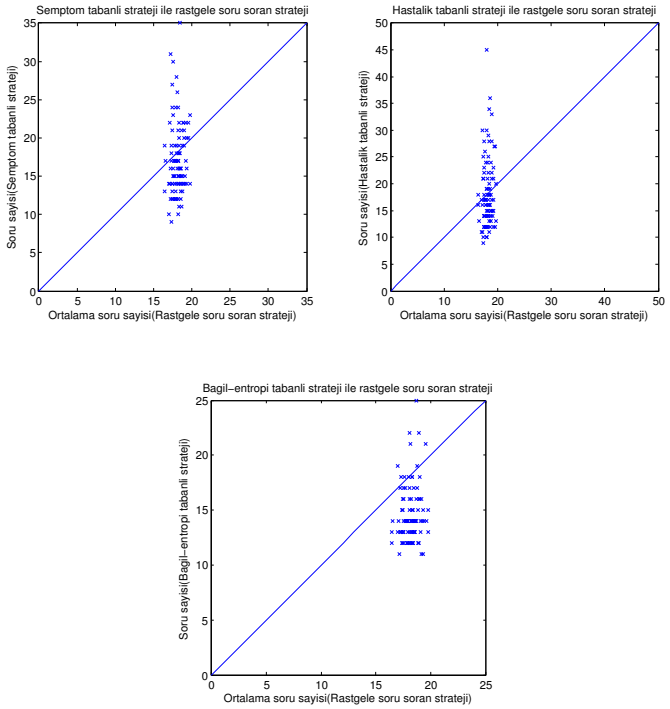
IV. Vargılar

Soru sıralama stratejilerinin başarımları güvenilir bir tanıya ulaşmak için sorulan soru sayılarının karşılaştırılmasıyla değerlendirildi. Stratejiler, soruyu yani hastaya sorulacak bir sonraki semptomu üretmektedir ve her adımda gözlemlenen semptomlara bağlı olarak hastalıkların sonsal dağılımlarının çıkarımını yapmaktadır. Bu sıralı tanı (sequential diagnosis) yaklaşımında, ard arda yapılan sorgulamaların sonuçlarına dayanarak çoklu tanı gerçekleştirildi. Soru sorma stratejileri ele alındığında, küçük ağ verisinde semptom tabanlı stratejinin rastgele soru soran stratejiden daha üstün olduğu görüldü. Diğer taraftan, büyük ağ verisinde semptom tabanlı stratejinin daha iyi olduğunu söylemek zor. Hastalık tabanlı strajinin hem küçük hem de büyük ağ verisinde rastgele soru soran stratejiden daha iyi olduğu söylenemez. Bağıl-entropi tabanlı stratejinin ise hem küçük hem de geniş ağ verisinde ortalama soru sayısı açısından açıkça rastgele soru soran stratejiden daha iyi olduğu görüldü. Bunun sebebi ise, bağıl-entropi tabanlı stratejinin hastanın sorulan sorulara verdiği yanıtlarının hesaba katmasıdır.

Hesaplama maliyetleri karşılaştırıldığında, bağıl-entropi tabanlı strateji semptom ve hastalık tabanlı stratejiye göre daha pahalıdır. Bağıl-entropi tabanlı stratejinin hesaplama maliyetinin yüksek olması da sorulacak bir sonraki soruyu bulurken yapılan çok sayıda çıkarımdan kaynaklanmaktadır. Buradaki hesaplama yükünü hafifletmek için, sayma (enumeration) ile yapılan tam çıkarım algoritmasının yanı sıra



Şekil 2: Küçük ağ verisinde soru sıralama stratejilerinin ikili karşılaştırılması



Şekil 3: Geniş ağ verisinde soru sıralama stratejilerinin ikili karşılaştırılması

quickscore [7] adında bir tam çıkarım algoritması ve yaklaşık bir çıkarım algoritması olan Gibbs örnekleyicisi kullanıldı. Sayma ile yapılan tam çıkarım algoritmasının hesaplama karmaşıklığı toplam hastalık sayısı ile üsseldir. Ancak, quickscore algoritmasının hesaplama karmaşıklığı pozitif semptom sayısı ile üsseldir [7]. Pratik durumlarda, pozitif semptomların sayısı toplam hastalık sayısında az olduğu için quickscore uygulanabilir bir çıkarım algoritmasıdır. Pozitif semptomların sayısının çok fazla olduğu durumlarda ise, çıkarım yapmak için Gibbs örnekleyicisi kullanılabilinir. Semptom tabanlı ve hastalık tabanlı stratejilerde ise soru sorma işlemi oldukça ucuzdur. Çünkü bu stratejiler ağın yapısına bağlıdır. Her iki strateji de hastanın sorulan sorulara verdiği cevaplardan bağımsız olarak sorulacak soruların permütasyonlarını üretir.

KAYNAKÇA

[1] Moshe Ben-bassat and D. Teeni, "Human-oriented information acquisition in sequential pattern classification: Part i x2014; single membership classification," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. SMC-14, no. 1, pp. 131–138, 1984.

[2] King-Sun Fu, *Sequential methods in pattern recognition and machine learning*, Mathematics in science and engineering. Academic Press, New York, 1968.

[3] M.A. Shwe, B. Middleton, D.E. Heckerman, M. Henrion, F.J. Horvitz, H.P. Lehmann, and G.E. Cooper, "Probabilistic diagnosis using a reformulation of the internist-1/qmr knowledge base. i. the probabilistic model and inference algorithms," *Methods of Information in Medicine*, vol. 30, pp. 241–255, 1991.

[4] Alice X. Zheng, Irina Rish, and Alina Beygelzimer, "Efficient test selection in active diagnosis via entropy approximation," in *Proceedings of 21st Conference on Uncertainty in Artificial Intelligence*, 2005, pp. 675–682.

[5] Gowtham Bellala, Jason Stanley, Clayton Scott, and Suresh K. Bhavnani, "Active diagnosis via auc maximization: An efficient approach for multiple fault identification in large scale, noisy networks," in *Proceedings of 27th Conference on Uncertainty in Artificial Intelligence*, 2011, pp. 35–42.

[6] E.J. Horvitz, D.E. Heckerman, B.N. Nathwani, and L.M. Fagan, "The use of a heuristic problem-solving hierarchy to facilitate the explanation of hypothesis-directed reasoning," in *In Proceedings of Medinfo*, 1986, pp. 27–31.

[7] David Heckerman, "A tractable inference algorithm for diagnosing multiple diseases," in *In UAI-89*, 1989, pp. 163–172.