Modellezés magasabb rendű Markov láncokkal

Egyed Tünde
Témavezető: Csiszár Villő

1. Bevezetés

A sztochasztikus folyamatok modellezése érdekes kérdéseket vet fel. Jó módszer lehet a modellezésükre, hogy a folyamatra Markov-láncként tekintünk, vagyis azt feltételezzük, hogy a jövőbeli érték, csak a jelenbeli értéktől függ, a múltbeli értékektől nem. Ez formálisan azt jelenti, hogy

$$P(X_{n+1} = x_{n+1}|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_{n+1} = x_{n+1}|X_n = x_n)$$

A dolgozatomban azt vizsgálom, hogy bizonyos esetekben nem lenne-e hasznosabb, ha a folyamatot magasabb rendű Markov-lánccal modelleznénk, azaz hosszabb memóriát feltételeznénk. Ennek azonban egyik hátránya, hogy a modell rendjének növelésével nő a modell komplexitása, ezáltal több adatra van szükség a megbízható modellezéshez, ami nem mindig áll rendelkezésre. A projektem célja bemutatni, hogy hogyan érdemes meghatározni egy Markov-lánc rendjét.

A kutatómunkám során két cikket tanulmányoztam a témában, melyek különböző módszerekkel, mutatókkal vizsgálták hálózatok rendjét.

A [1]-es cikkben az internetes oldalakon való böngészést figyelték meg. Itt a kutatás célja a hálózat emlékezetének meghatározása volt. Mivel a felhasználók linkeken keresztül közlekednek, ezért azt várjuk, hogy a legalább elsőrendű modellek jobban teljesítenek, mint a nullarendű modell, azonban az eredményekből az látszik, hogy weboldalak szintjén a nullarendű modell mutatói bizonyultak a legjobbnak. Ennek oka valószínűleg az, hogy a komplexebb modellek megtalálásához nem áll rendelkezésre elegendő adat. De ha a weboldalakat témakörök szerint csoportosítjuk, akkor már a mutatókban is megfigyelhető a magasabb rendű modellek eredményessége.

A [2]-es cikkben különböző hálózatokon vizsgálták az első- és a másodrendű modellek közti különbséget.

2. Modellezés

A félév során szimulált adatokon próbáltam ki a szakirodalomban bemutatott módszerek közül kettőt. Első lépésben egy másodrendű mintát szimuláltam. A szimuláció során 20 darab 100 elemű mintát készítettem. Ezek négy különböző lehetséges értéket vehetnek fel ("A", "B", "C", "D"). Ahhoz, hogy elérjem a másodrendű tulajdonságot,

az n. elem véletlen generálásánál az $n-1$. elemet négyszeres, az $n-2$. elemet háromszoros súllyal, a többi elemet egyszeres súllyal vettem figyelmbe. A szimulált adatokat első- és másodrendű Markov-lánccal modelleztem, majd azt vizsgáltam, hogy mennyivel fest részletesebb képet a másodrendű modell az elsőrendű modellhez képest.

Ehhez először a [1]-ben is ismertetett loglikelihood módszert alkalmaztam. A likelihood függvény felírásához megbecsültem a minta alapján az első és másodrendű átmenetmátrixokat, ahol az átmenetvalószínűségek becsléseként a relatív gyakoriságot vettem. Ez a módszer a másodrendű modell esetén úgy alkalmazható, hogy a modellt visszavezetjük elsőrendűre, vagyis például az $A \to B \to A$ átmenetre $AB \to A$ átmenetként tekintünk. A likelihood függvény az átmenetek valószínűségeinek szorzata:

$$L(\theta_k, x) = p(x_n|x_{n-1})p(x_{n-1}|x_{n-2}) \ldots, p(x_2|x_1)p(x_1) = p(x_1) \prod_i \prod_j p_{ij}^{n_{ij}}$$

ahol θ_k a k-ad rendű modell átmenetmátrixa, p_{ij} annak a becsült valószínűsége, hogy x_i-ből x_j-be megyünk, n_{ij} a mintában az átmenetek száma x_i-ből x_j-be.

A várakozásaimnak megfelelően a másodrendű modell esetén magasabb loglikelihood érték adódott. Önmagában ez még nem elég ahoz, hogy azt mondhassuk, hogy a másodrendű modell jobb. Ehhez vizsgálnunk kell, hogy a két loglikelihood közti különbség szignifikáns-e. Ennek eldöntésére valószínűséghányados próbát alkalmaztam. Ehhez null modellként $k = 1$ rendet feltételeztem, alternatív modellként $m = 2$ rendet, a próbastatisztikát pedig a következő képlettel kapom:

$${}_k\eta_m = -2(logL(\theta_k, x) - logL(\theta_m, x))$$

Az [1] szerint a próbastatisztika χ^2 eloszlású $(|S|^m - |S|^k)(|S| - 1)$ szabadságfokkal, ahol $|S|$ a lehetséges állapotok száma, esetünkben $|S| = 4$. Ezzel a módszerrel a mintán p értékként nullát kaptam, vagyis a másodrendű modell valóban jobbnak bizonyult.

Második módszerként a [2]-ben bemutatott kétlépéses módszert alkalmaztam. Itt először megbecsültem, hogy mennyi a valószínűsége annak, hogy egy adott állapotba két lépés múlva ismét visszatérek. Ehhez a becsült átmenetmátrix alapján felírjuk $P(X_{n+2} = i|X_n = i, X_{n+1} = j)$ valószínűséget minden i-re. Ezután ezen értékek összegét vettem súlyozva a stacionárius eloszlással. Az így kapott első- és másodrendű valószínűségeket összevetettem a mintában előforduló kétlépéses visszatérések relatív gyakoriságával. Végeredményben azt tapasztaltam, hogy a másodrendű modellből kapott érték közelebb áll a relatív gyakorisághoz, mint az elsőrendű modell. Tehát a másodrendű modell ebből a szempontból is jobban teljesít.

Hivatkozások

[1] Singer, Philipp, et al. "Detecting memory and structure in human navigation patterns using markov chain models of varying order." PloS one 9.7 (2014): e102070.

[2] Rosvall, Martin, et al. "Memory in network flows and its effects on community detection, ranking, and spreading." Ecology 19 (2014): 30.