Comparação de assinaturas de amostras em Árvores Probabilísticas de Contexto

Marina Lobato, Diego Leal e Denise Duarte
Depto de Estatística - ICEX- UFMG

29 de abril de 2010

1 Resumo

Introduzidas por Rissanem em 1983, as árvores probabilísticas de contexto (PCT) são uma classe promissora de modelos que podem auxiliar na área da genética, lingüística ou qualquer outra, onde as amostras sejam sequencias de dados discretos e se tenha interesse em encontrar um modelo gerador para os dados. As PCT também são conhecidas na literatura como Variable Length Markov Chains (VLMC). Em contraste com os modelos de cadeia de Markov, onde cada variável no tempo t depende de um número fixo de variáveis no passado, em modelos de PCT, o tamanho do passado relevante para prever o próximo simbolo pode variar com base na realização específica observada.

Observe na tabela abaixo, considerando como exemplo uma cadeia de Markov onde temos somente 2 estados , como o número de parâmetros livres cresce quando aumentamos sua ordem.

Ordem	0	1	2	3	4	5	6	7	8	9	10
N^o	2	6	18	54	162	486	1458	4374	13122	39366	118098

Em um modelo de cadeia de Markov univariado convencional o número de parâmetros do modelo cresce exponencialmente com a ordem da cadeia, $O(|A|^n)$, onde $|A|$ é o tamanho do alfabeto de n é a ordem da cadeia. Isto é desanimador para qualquer um que pretenda usar este modelo para estimar a lei do processo.

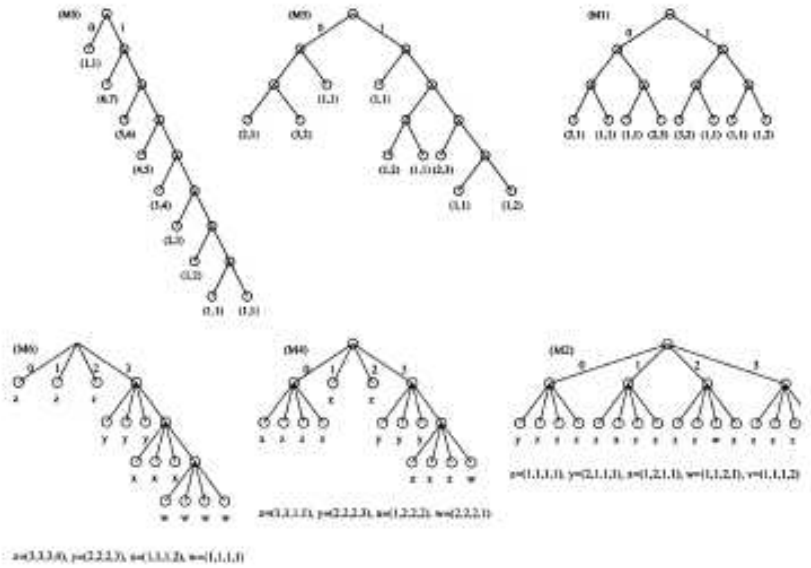
De um ponto de vista prático, não é interessante um modelo que considere todas as seqüências de tamanho fixado,k, possíveis porque teríamos que estimar um número muito grande de parâmetros.

As árvores probabilísticas definem uma família de cadeias estocásticas de ordem finita (ou mesmo infinita) em um alfabeto finito. A idéia é que para cada passado, apenas um sufixo finito do passado (sequência finita de símbolos), chamada de *contexto* é suficiente para predizer o próximo símbolo. Esses sufixos podem ser representados por uma árvore enumerável completa de contextos finitos. Este modelo é bem mais flexível e parcimonioso do que cadeias de Markov de alcance fixo.

Em uma árvore probabilística de contextos existe uma probabilidade de transição associada a cada contexto. Uma cadeia estacionária é consistente com uma árvore probabilística de contextos se para todo passado infinito e qualquer símbolo temos que as seguintes condições são satisfeitas

1. Para nenhum $w_{-1}^{-k} \in \tau$ temos que $w_{-1}^{-k+1} \in \tau$ para $j = 1, ..., k$

2. Nenhuma sequencia pertencente a τ pode ser substituída por um contexto sem violar a propriedade descrita no item anterior onde w_{-1}^{-k} é o único elemento da árvore que é um contexto da sequência.



Então seja $p = \{p(.|w) : w \in \tau\}$ uma família de medidas de probabilidade em um alfabeto A indexada pelo elementos de τ. O par (τ, p) é chamado de *probabilistic context tree* (PCT).

1.1 Estimação da PCT dada uma amostra

Alguns algoritmos tem sido apresentadas na literatura para estimar uma árvore probabilística de contexto(PCT) e também as probabilidades de transição. Uma lista incompleta inclui Ron et al. (1996), Buhlmann e Wyner (1999)(VLMC) e Galves et al. (2009). Uma abordagem diferente é proposta Csiszar e Talata (2006). Eles mostraram que uma árvore probabilística de contexto pode ser consistentemente estimada em um tempo linear usando o Critério de Informação Bayesiana (BIC)(1995).

Em Galves, Galves, Garcia e Leonardi (2009) é introduzido o critério do menor maximizador (smallest maximizer criterion) para estimar uma PCT. Este critério seleciona a árvore na classe das campeãs estimadas pelo BIC, para cada valor da constante de penalização. Este algoritmo é chamado de G3L.

A constante de penalização pode ser interpretada como o valor que se "paga"por estimar um grande número de parâmetros. O estimador da árvore probabilística de contexto com constante de penalização $c > 0$ é difinido como:

$$\hat{\tau}_{bic(X_1^n;c)} = argmax_{\tau \in Tn} log L_\tau(X_1^n) - c\Delta df(\tau)\Delta logn \tag{1}$$

Onde $L_\tau(X_1^n)$ é a verossimilhança da árvore τ dada a amostra e $df(\tau)$ denota o número de graus de liberdade do modelo correspondente da árvore de contexto τ. Com o subconjunto das árvores obtidos pela variação da constante de penalização no critério BIC, que garante um bom modelo para os dados históricos, mas penaliza a complexidade do mesmo. Com o objetivo de selecionar o modelo, o algoritmo G3L escolhe o conjunto de árvores campeãs Cn contido em τ. Isso é feito explorando a relação entre o menor critério maximizador e o BIC.

Csiszar e Talata(2006) provam a consistência do procedimento da seleção BIC no caso de árvores ilimitadas quando $d(n) = o(logn)$. Junto da consistência desse procedimento, essa condição também implica que a estimação pode ser feita em um tempo linear, usando o maximizador da árvore de contexto(CTM), um algoritmo introduzido por Willems et al.(1995).

Galves et al(2009) mostram que todas as árvores campeãs $\mathcal{C}_n$ podem ser obtidas usando o algoritmo CTM pela mudança da constante de penalização do BIC. O algoritmo G3L nos dá a árvore campeã entre os candidatos a árvore dada pela amostra utilizando o critério BIC de seleção. Mas isso também dá uma sequência de árvores geradas pela amostra de cada função da constante de penalização. Essa função pode ser vista como uma evolução da amostra de acordo com o preço que se paga para obter o modelo mais simples para os dados. Por exemplo, se escolhemos um modelo independente, ordem zero, quando a ordem verdadeira é grande, então a constante de penalização deverá ser bem grande.

1.2 Assinatura de uma amostra

Dada uma amostra $X_1, X_2, X_3, ..., X_n$, a árvore campeã estimada pelo $G3L$, $\hat{\tau}_g$ é uma função da amostra e a constante de penalização ótima c_{opt},

$$\hat{\tau}_g = f(X_1^n, c_{opt}) \tag{2}$$

Essa constante ótima é obtida mudando os valores da constante de penalização e escolhendo aquela com o maior valor do BIC. Mas para cada valor de c, temos um valor da penalização da verossimilhança. Desse modo, antes de estimar um valor ótimo para c (segundo os critérios propostos por Galves et al), o algoritmo G3L gera uma sequencia de constantes de penalização.

$$0 < c_1 < c_2 < \ldots < c_{opt} \tag{3}$$

Que nos leva a uma sequencia de árvores

$$\tau_0 \succ \tau_1 \succ \tau_2 \succ \ldots \succ \tau_{opt} \tag{4}$$

onde o símbolo $\succ$ significa que uma árvore é maior do que a outra em número de galhos. Chamamos a sequência $C_x = (0, c_1, c_2, \ldots, c_{opt})$ de **Assinatura da amostra**.

Dada uma amostra $X_1, \ldots, X_n$, C_x é uma sequencia de constantes de penalização que caracteriza esta amostra no seguinte sentido:

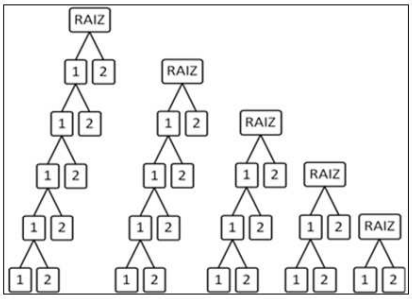
1. Em amostras da mesma PCT, as sequencias de constantes de penalização não diferem muito umas das outras.

2. Em amostras de duas árvores diferentes as sequencias de constantes de penalização são diferentes.

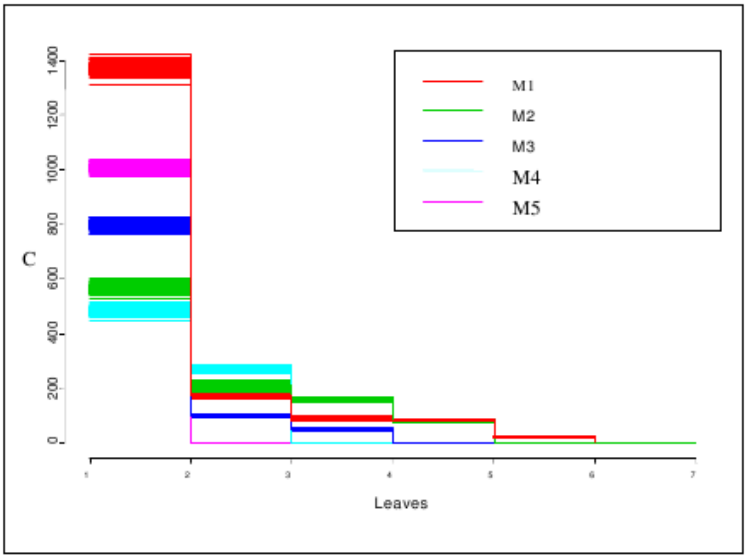
1.3 Resultados de simulação

Através de estudos de simulação observou-se que se X_1^n e Y_1^n são duas amostras da mesma árvore de contexto probabilística e C_x e C_y são as respectivas assinaturas das amostras , então para um valor de n suficientemente grande, se as assinaturas da amostra são diferentes, elas não provêm da mesma árvore de contexto probabilística.

1.4 Modelos simulados M1, M2, M3, M4, M5



1.4.1 Assinaturas dos modelos M1, M2, M3, M4, M5 para 500 amostras



1.5 Teste de hipóteses bootstrap para comparação de assinaturas

Queremos propor um teste bootstrap, baseado na assinatura da amostra, para decidir se duas amostras foram geradas a partir da mesma árvore de contexto probabilística. As etapas do teste são as seguintes:

1. Dada uma amostra $X_1, ..., X_n$, estimamos a árvore $\hat{\tau}_X$.
2. A partir da árvore $\hat{\tau}_X$, geramos B reamostras de tamanho n $X^*_{1,1}, ..., X^*_{1,n}$, ..., $X^*_{B,1}, ..., X^*_{B,n}$.
3. Para cada reamostra calculamos sua assinatura , obtendo então B assinaturas $C_{X^*_1}, \ldots, C_{X^*_B}$.
4. Construímos uma banda de confiança bootstrap tomando os percentis 5% e 95% dos valores de c para cada árvore τ associada. Chamamos a sequencia dos percentis 5% de $C_{X_{0.05}}$ e a sequencia dos percentis 95% de $C_{X_{0.95}}$.
5. Dada uma amostra $Y_1, ..., Y_n$ estimamos sua assinatura C_Y.
6. Fazemos o gráfico de $C_{X_{0.05}}$, $C_{X_{0.95}}$ e C_Y contra o número de folhas da árvore.
7. Se C_Y estiver contida na banda de confiança bootstrap entre $C_{X_{0.05}}$, $C_{X_{0.95}}$, não rejeitamos a hipótese de que $X_1, ..., X_n$ e $Y_1, ..., Y_n$ sejam amostras da mesma árvore. Caso contrário, concluímos que as amostras vêm de árvores diferentes.

Apoio FAPEMIG

Referências

[1] Bühlmann (2000) Model selection for variable length Markov chains and tun- ing the context algorithm. Ann. Inst. Statist. Math., 52(2).

[2] Bühlmann and A. J. Wyner (1999) Variable length Markov chains. Ann. Statist., 27, 1999.

[3] Csiszar, I. and Talata, Z. (2006) Context tree estimation for not necessarily nite memory processes, via BIC and MDL. IEEE Trans. Inform. Theory, 52(3)

[4] Duarte, D., Galves, A. e Garcia, N. L. (2006) Markov approximation and consistent estimation of unbounded probabilistic suffix trees. Bulletin of the Brazilian Mathematical Society.

[5] Galves, A., Galves, C., Garcia, N. and Leonardi, F. (2009) Context tree selection and linguistic rhythm retrieval from written texts. ArXiv: 0902.3619v2.

[6] Raftery, A. (1985) A model for high-order Markov Chains. Journal of the Royal Statistical Society B, 47, 528-539.

[7] Rissanen, J.(1983) A universal data compression system. IEEE Trans. Inform.Theory, 29(5).

[8] Willems F.M.J., Shtarkov Y.M., and Tjalkens T.J.(1995) The context tree weighting method: Basic properties. IEEE Trans. Inf. Theory, 41,653-664.