COEFICIENTE DE SIMILARIDADE DA MEDIDA INVARIANTE DO RITMO DE TEXTOS HISTÓRICOS BRASILEIROS

Silvio Alves de Souza[1], **Denise Duarte**[2], **Eduardo M. A. M. Mendes**[3]

Resumo: Neste trabalho introduzimos o conceito *similaridade* entre duas amostras de árvores de sufixo probabilísticas (PST), com espaço de estados discreto. Pretendemos identificar o quanto de similiridade existe entre estas cadeias atraves de suas medidas invariantes. Trabalharmos com a PST com o objetivo de contornar o problema de quantidade de parâmetros a serem estimados em modelos de cadeias de Markov de ordem fixa.
Palavras-chave: *Cadeias de Markov; PST; PSA; Coeficiente de similaridade.*

1 Introdução

Nossa motivação foi dada por [1] que apresenta uma forma recurssiva para determinar a distribuição invariante de determinada cadeia como combinação linear de outras. Assim adaptamos este método para o caso de Probabilistic Suffix Tree (PST). A diferença entre nosso trabalho e o do [1] é que ele utiliza cadeias de Markov enquanto que nós utilizamos PST.

Optamos em utilizar PST na modelagem de sequencias com memoria, caso discreto, pois assim contornamos o problema do aumento do número de parâmentros da cadeia quando a ordem de dependência no passado não é pequena. Neste caso, o número de parâmetros pode ser muito grande e não se pode estimá-los com um tamanho de amostra fixo. Veja, por exemplo, [2, 3, 5, 4].

2 Objetivo

O objetivo deste trabalho é encontrar os coeficientes de similaridade entre cadeias de textos ritmicos utilizando árvore de sufixo probabilístico (PSA).

3 Metodologia

3.1 Árvore de sufixo probabilística

Introduzido por [6], Probabilistic Suffix Trees (PST), é uma classe de cadeias estocásticas com memória de comprimento variável. Ao contrário dos modelos de cadeias de Markov, onde cada variável no tempo t depende de um número fixo de variáveis no passado, elas tem a propriedade de que, para cada string de simbolos passados, somente um sufixo infinito do passado é suficiente para predizer o próximo símbolo. Esta parte relevante do passado é denominada *contexto*.

Um subconjunto infinito τ de $\bigcup_{k=1}^{\infty} A^{\{-k,\ldots,-1\}}$ é uma árvore irredutível se ela satisfaz as seguintes condições:

[1]DECOM-Centro Federal de Educação Tecnológica de Minas Gerais, silvio@decom.cefetmg.br
[2]DEST-Universidade Federal de Minas Gerais, dduarte.est@gmail.com
[3]DELT-Universidade Federal de Minas Gerais, emmendes@cpdee.ufmg.br

Propriedade sufixo. Para nenhum $w_{-1}^{-k} \in \tau$, $w_{-1}^{-k+1} \in \tau$ para $j = 1, ..., k$.
Irredutibilidade. Nenhuma sequencia pertencente a τ pode ser substituido por um sufixo adequado sem violar a propriedade sufixo.

Este tipo de modelo é mais parcimonioso, isto é, ele tem menos parâmetros, $|\tau|(|A| - 1)$ ao invés de $|A|^k(|A| - 1)$ e $|\tau| << |A^k|$ em geral, onde k é a ordem da cadeia.

3.2 Automata finito probabilistico

Uma *Probabilistic Finite Automata* (PFA) é uma $5-tuple$ $(A, E, \alpha, \gamma, \pi)$, onde A é um *alfabeto*, E é um conjunto finito de strings composta por simbolos do alfabeto, $\alpha : E\ x\ A \mapsto\ E$ é a *função de transição* , $\gamma : ExA \mapsto [0, 1]$ é o *função de probabilidade do próximo simbolo* e $\pi : E \mapsto [0, 1]$ é a *distribuição inicial* sobre os estados de inicio de E. A função γ e π satisfazem as seguintes condições: para cada $e \in E$ $\sum_{a\in A} \gamma(e, a) = 1$ e $\sum_{e\in E} \pi(q) = 1$.

3.3 Automata sufixo probabilistico

Para qualquer inteiro N, seja A^N denotando todas as strings de comprimento N e $A^{\leq N}$ denotando o conjunto das strings com comprimento no máximo N. A Probabilistic Suffix Automata (PSA) é uma subclasse de um Probabilistic Finite Automata (PFA) de modo que cada estado de E em uma PSA é definido por uma string de comprimento finito em $\mathcal{A}$ e para cada dois estados e_1 e $e_2 \in E$ e para cada símbolo $a \in A$, se $\alpha(e_1, a) = e_2$ então $e_2 = e_1a$.

Para qualquer $L \geq 0$ a subclasse de PSA's em que cada estado é nomeado por uma string de comprimento no máximo L é denotado por $L - PSA$. Quando E inclue todas as strings em A^L isto é uma cadeia de Markov de ordem L nomeada $L_{full} - PSA$.

4 Emulação de uma PST por uma PSA

A equivalência entre uma PSA e uma PST apresentada em [7] é indicado nos seguintes teoremas.

Teorema 5.1: Para cada $L - PSA = (E, A, \alpha, \gamma, \pi)$ existe uma PST equivalente τ de profundidade máxima L e no máximo $L.|E|$ nós.

Teorema 5.2: Para cada τ de profundidade L sobre A, existe uma PFA equivalente M_τ, com no máximo $L.|T|$ estados. Além disso, se a propriedade é válida para τ, então existe uma PSA equivalente para cada string s nomeando um nó na árvore τ, $P_\tau = \sum_{a\in A} P_\tau(as)$.

5 Cadeia de Markov multivariada

[1] propõe uma metodologia para lidar com modelo de cadeia de Markov multivariada. Eles queriam modelar sequências multivariadas de dados categóricos gerados por fontes similares. Considere sequências de m Cadeias de Markov sobre A cada uma com matriz de probabilidade de transição Q_{a_i,a_j}, $a_i, a_j \in A$ e $Q^{(hk)_{a_i,a_j}}$ é a matriz de probabilidade de transição para os estados na h^{th} sequências para os estados na k^{th} sequência. Para um modelo de Markov de primeira ordem temos a seguinte relação :$\pi^{\mathbf{j}}{}_{n+1} = \sum_{k=1}^{m} \lambda_{jk} Q^{(jk)} \pi_n^k$ para $j = 1, 2, ..., s$ e $n = 0, 1, \ldots$ onde $\lambda_{jk} \geq 0$, $1 \leq j, k \leq s$ e $\sum_{k=1}^{s} \lambda_{jk} = 1$, para $j = 1, 2, ..., s$ e $\pi_{\mathbf{0}}^{\mathbf{j}}$ é a distribuição de probabilidade inicial da $j - th$ sequência.

6 PST Multivariada

No caso de nao haver uma PSA completa equivalente a um PST , também seria possível pensar de um PST como uma cadeia de Markov de ordem L. Para este fim, é necessário para completar a árvore com os ramos ausentes até a ordem L que tem a mesma probabilidade de transição que o seu pai. Desta forma a cadeia de Markov de ordem um equivalente terá novamente $|A|^L$ estados compostos por todas as strings de tamanho L com matriz de transição $Q_L(x, y)$.

Basta considerar um PST como PSA baseado no fato de que os teoremas 5.1 e 5.2 garantem que há uma equivalência entre eles.

Seja λ os parâmetros peso estimados usando a verdadeira matriz de transição $Q_\tau^{(jk)}$ e seja $\hat{\lambda}$ os parâmetros peso estimados usando a matriz de transição estimada $\hat{Q_\tau}^{(jk)}$. Então $\lambda \to \hat{\lambda}$ em probabilidade.

7 Resultados e discussões

O Tycho Brahe Parsed Corpus of Historical Portuguese é um corpus eletrônico de textos escritos em Português por autores nascidos entre 1380 e 1845. O corpus é encontrado em (http://www.tycho.iel.unicamp.br/ tycho). O objetivo é comparar alguns desses textos por estimação da similaridade de acordo com as características rítmicas de cada texto utilizando a metodologia proposta. Textos escritos pelo mesmo autor, autores nascidos ao redor do mesmo período de tempo e de textos escritos por autores nascidos em um período de tempo diferente são comparados. Mostramos como cada sílaba foi codificada: 0 (sílaba átona no meio ou final de palavras), 1 (sílaba tônica no meio ou final de palavras), 2 (sílaba átona no início de palavras), 3 (sílaba tônica no início de palavras), 4 (fim de sentença).

7.1 Estimando os coeficientes de similaridade dos textos

O principal objetivo é estimar a similaridade entre os textos escritos pelo mesmo autor ou autores diferentes em diferentes períodos de tempo ou não. Para esse fim, textos de Almeida Garret (1799), textos de Antonio Vieira (1608), textos escritos não muito distante por diferentes autores, Marquês D'Alorna (1802) e Marquesa D'Alorna (1750) e, finalmente, textos escritos muito tempo separados por diferentes autores, Gandavo (1502) e Ramalho Urtigão (1836), são comparados.

Considere: π^1 = medida invariante do texto 1; π^2 = medida invariante do texto 2; P_{11} = matriz de probabilidade de transição do texto 1 para ele mesmo; P_{12} = matriz de probabilidade de transiçao cruzada do texto 2 para o texto 1; P_{21} = matriz de probabilidade de transiçao cruzada do texto 1 para o texto 2; P_{22} = matriz de probabilidade de transiçao do texto 2 para ele mesmo;

7.1.1 Coeficiente de similaridade entre textos escritos pelo mesmo autor

1. Similaridade entre dois textos escritos por Almeida Garret, 1799 (G003 e G004).

   Aplicando a metodologia proposta temos:$\pi_{n+1}^{G003} = 0.6717P^{(11)}\pi_n^{G003} + 0.3283P^{(12)}\pi_n^{G004}$ e $\pi_{n+1}^{G004} = P^{(22)}\pi_n^{G004}$

   Note que a medida invariante do texto G003 depende da medida invariante de ambos os textos. Isto sugere que estes textos tem alguma similaridade. Contudo a medida invariante de G004 depende somente dela mesma.

2. Coeficiente de similaridade entre dois textos escritos por Antonio Vieira, 1608(V002 e V004).

   Neste caso encontramos: $\pi_{n+1}^{V002} = 0.7568P^{(11)}\pi_n^{V002} + 0.2432P^{(12)}\pi_n^{V004}$ e $\pi_{n+1}^{V004} = P^{(22)}\pi_n^{V004}$

   Observe que o texto V004 depende somente dele mesmo, enquanto o texto V002 tem alguma similaridade com o texto V004.

7.1.2 Autores nascidos no mesmo período de tempo

Nesta seção estimamos a similaridade entre os textos escritos por Marquês D'Alorna (A003), nascida em 1802 e Marquesa D'Alorna (A004) nascida em 1750.

A relaçao encontrada foi $\pi_{n+1}^{A003} = 0.9521P^{(11)}\pi_n^{A003} + 0.0479P^{(12)}\pi_n^{A004}$ e $\pi_{n+1}^{A004} = 0.0013P^{(21)}\pi_n^{A003} + 0.9987P^{(22)}\pi_n^{A004}$

Note que não há pouca similaridade entre estes textos.

8 Conclusões

Percebemos que os coeficientes de similaridade nos dão uma boa idéia da similaridades entre os textos. Isto mostra o quanto os ritmos estão relacionados.

Referências

[1] Ching, W. K., A multivariate Markov chain model for categorical data sequences and its applications in demand predictions, *Journal of Management Mathematics*. **13**, 187-199, 2002.

[2] Bejerano, G., Algorithms for variable length Markov chain modeling, *Bioinformatics*, **20**, 788-789, 2004.

[3] Browing, S. R., Multilocus association mapping using variable-length Markov chains, *The American Journal of Human Genetics*, **78(6)**, 903-913, 2006.

[4] Buhlmann, P., Wyner, A. J., Variable length markov chains, *The Annals of Statistics*, **27(2)**, 480-513, 1999.

[5] Leonardi, F. G., A generalization of the PST algorithm: modeling the sparse nature of protein sequences, *Bioinformatics Advanced*, **22**, 1302-1307, 2006.

[6] Rissanen, J., A Universal Prior for Integers and Estimation by Minimum Description Length, *The Annals of Statistics*, **11(2)**, 416-431, 1983.

[7] Ron, D., Singer, Y., Tishby, N., The power of amnesia: Learning probabilistic automata with variable memory length, *Machine Learning*, **25**, 117-149, 1996.