Aprendizado Baseado na Teoria da Informação: Fundamentos e Perspectivas

Daniel Guerreiro e Silva , Romis Attux

Departamento de Engenharia de Computação e Automação Industrial (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Caixa Postal 6101, 13083-970 – Campinas, SP, Brasil

{danielgs,attux}@dca.fee.unicamp.br

Abstract - This paper briefly introduces a new research field that is called Information Theoretic Learning, which is based on statistics that are more informative than the second-order statistics adopted by traditional adaptive algorithms. Besides exposing the motivation and some definitions, we present some application examples and investigation opportunities.

Keywords: machine learning, information theory, information theoretic learning, adaptive algorithms.

1. Introdução

Algoritmos de aprendizado se caracterizam por realizarem o ajuste de parâmetros de um modelo através da otimização de um critério que indique seu desempenho frente aos dados apresentados. Ao longo dos anos, um critério que vem sendo largamente utilizado para essa tarefa baseia-se em estatísticas de segunda ordem do erro entre o sinal de saída do mapeador e um sinal de referência.

Há diversas razões para o uso de um critério baseado no segundo momento dos dados. Entre elas, podem-se destacar [3, 14]: (i) é simples de usar; (ii) possui o significado físico de ser uma medida de energia do sinal em questão; (iii) é um critério com propriedades interessantes no contexto de otimização, como diferenciabilidade e simetria; (iv) é muito bem-sucedido na solução de problemas pertencentes ao domínio linear-gaussiano e (v) origina uma enorme variedade de algoritmos adaptativos.

Por outro lado, sabendo que é ideal extrair o máximo de informação dos dados durante a adaptação dos parâmetros, há evidências que indicam que o segundo momento é uma medida pobre para essa tarefa de avaliar a equivalência de informação entre o sinal desejado e a saída do mapeador [7]. Além disso, o avanço da capacidade computacional e o estudo de problemas mais complexos em processamento de sinais levam-nos a cenários onde esta tradicional família de critérios pode não ser a mais satisfatória.

A Teoria da Informação (TI), desenvolvida a partir de 1948 por Claude E. Shannon [10], lida com a quantificação da incerteza e da dependência estatística em processos aleatórios, ao mesmo tempo que vincula tais medidas ao conceito de informação. Esta área do conhecimento contribuiu em parte com o enorme desenvolvimento dos sistemas de comunicação daquela época até hoje.

Através do trabalho pioneiro de Principe et al., de 2000, que define o Aprendizado Baseado na Teoria da Informação ou *Information Theoretic Learning (ITL)* [7], surge então no estudo dos algoritmos adaptativos o interesse pelo uso de critérios derivados a partir de TI e que permitiriam superar as limitações das estatísticas de segunda ordem. Tal progresso pode ser obtido tanto no universo de dados contínuos, sobre o qual esta nova área se iniciou, como no universo de dados discretos, que mesmo com suas peculiaridades próprias já dispõe de esforços iniciais para uso destes novos critérios na solução de problemas. Nas próximas seções, resumem-se brevemente os passos dados pela pesquisa em ITL até a atualidade e busca-se apontar caminhos para novas contribuições, as quais estão sendo alvo do trabalho de doutorado do autor.

2. Entropia e Informação Mútua

Entropia é o conceito primordial no estudo de TI e indica o grau de incerteza médio associado a uma determinada variável aleatória (VA) discreta:

$$H(X) = -\sum_{x \in \chi} p(x) \log[p(x)], \qquad (1)$$

onde $p(x)$ é a função massa de probabilidade e χ é o conjunto de possíveis valores assumidos pela VA X.

A extensão desta definição para o caso contínuo é denominada entropia diferencial, e seu cálculo para uma VA X é:

$$h(X) = -\int f(x)\ln[f(x)]dx, \qquad (2)$$

onde $f(x)$ é a função densidade de probabilidade de X. Outro conceito fundamental é o de Informação Mútua entre duas VAs, X e Y:

$$I(X;Y) = \sum_{x\in\chi}\sum_{y\in\Upsilon} p(x,y)\log\left[\frac{p(x,y)}{p(x)p(y)}\right], \quad (3)$$

onde $p(x,y)$ é a função massa de probabilidade conjunta, $p(x), p(y)$ são as funções de probabilidade marginais e Υ é conjunto de possíveis valores assumidos pela VA Y. Também há a extensão do conceito para o caso de variáveis aleatórias contínuas:

$$I(X;Y) = \int\int f(x,y)\ln\left[\frac{f(x,y)}{f(x)f(y)}\right]dxdy, \qquad (4)$$

onde $f(x,y)$ é a função densidade de probabilidade conjunta e $f(x), f(y)$ são as funções de densidade marginais. A entropia pode ser vista como uma generalização da variância, enquanto a informação mútua é uma medida de independência entre as variáveis, generalizando o conceito de correlação [3]: se duas variáveis possuem informação mútua nula, então elas são estatisticamente independentes.

Dadas estas definições, ITL é a otimização não-paramétrica de sistemas adaptativos através do uso de critérios de desempenho baseados em TI, como a Entropia, a Informação Mútua e outros [1]. Embora a área tenha sido concebida, inicialmente, para tratar do cenário de dados contínuos e com algoritmos de adaptação baseados em derivadas da função custo, propomos a ampliação do escopo de ITL para dados discretos e para outras abordagens (por exemplo, meta-heurísticas evolutivas) de adaptação dos parâmetros quando o cálculo da derivada é inviável.

3. Aplicações de ITL

3.1. Dados contínuos

Há diversas formulações de critérios baseados em TI para solucionar problemas de aprendizado supervisionado e não-supervisionado. Para o primeiro caso considere, por exemplo, uma máquina que faz o mapeamento $f(\mathbf{x},\mathbf{w}) = y$ de dados de um vetor entrada $\mathbf{x}$ para uma saída y e que tem o conjunto de parâmetros $\mathbf{w}$ ajustados de tal forma que y se "aproxime" ao máximo da saída d desejada, o que, no paradigma de TI, é tentar aproximar a distribuição conjunta $f_{\mathbf{w}}(\mathbf{x},d)$ da distribuição $f(\mathbf{x},d)$. É possível demonstrar que, no contexto de identificação de sistemas, isto ocorre se for solucionado o problema de minimizar a entropia do sinal de erro $e = d - y$ [2]:

$$min_{\mathbf{w}} \quad h(E) = -\int f_{\mathbf{w}}(e)\ln[f_{\mathbf{w}}(e)]de. \qquad (5)$$

Já em um problema de aprendizado não-supervisionado, pode-se utilizar o princípio de máxima transferência de informação (InfoMax), que consiste em maximizar com respeito a $\mathbf{w}$ a informação mútua entre o sinal de entrada do mapeador ($\mathbf{x}$) e o sinal de saída (y) [3]. Outra abordagem, utilizada no contexto de análise de componentes independentes, é a de minimizar a informação mútua entre os componentes da saída do modelo.

Identificação de sistemas não-lineares [2], separação cega de fontes [5], extração de características [12] e clusterização [6] são exemplos de problemas com aplicação de algoritmos baseados nas formulações apresentadas e em outras formulações de critérios baseados em TI.

3.2. Dados discretos

Pela pesquisa realizada na literatura até agora, notamos que ITL tem se relacionado fundamentalmente a problemas modelados por dados com valores reais. Entretanto, recentes trabalhos [4, 15, 16] discutem a possibilidade de se realizar análise de componentes independentes em sinais com alfabeto finito.

Aplicando esta ideia ao contexto de separação cega de fontes, o cenário compreende um conjunto de fontes independentes de alfabeto finito que sofre uma transformação linear no mesmo domínio, ou seja, produzem-se misturas com o mesmo alfabeto.

Tanto Gutch et al. como Yeredor propõem critérios baseados na minimização da entropia dos sinais extraídos a fim de obter as componentes independentes, por isso podem ser considerados no

escopo de ITL. Ainda que as contribuições estejam atualmente restritas ao campo teórico, é viável investigar possibilidades de aplicação em codificação, análise de fatores, mineração de dados gênicos, entre outros. Uma outra oportunidade em aberto e que estamos investigando é o uso de algoritmos evolutivos associado ao critério de minimização da informação mútua para determinar a matriz de separação, e assim obter de uma só vez todos os componentes independentes.

4. Estimadores

Dado que, nos problemas de aprendizado de máquina, há uma amostra finita de dados para treinamento e geralmente não se conhece sua distribuição, uma questão crucial para derivar o algoritmo de adaptação em ITL é que se utilizem estimadores das distribuições e da entropia (ou outra medida associada a TI), tanto no caso contínuo como no caso discreto.

4.1. Caso contínuo

Os principais trabalhos nesse caso utilizam o método de janela de Parzen para estimar a densidade de probabilidade dos dados, o qual consiste de aproximar a distribuição por uma soma de funções *Kernel* centradas nas amostras. Entre as vantagens desse método está o fato de se poder escolher um *Kernel* contínuo e diferenciável, viabilizando métodos de adaptação por gradiente. Por outro lado, o número de amostras disponíveis e a escolha adequada dos parâmetros do *Kernel* são fatores sensíveis e assim podem prejudicar o desempenho na estimação, se mal considerados.

Quanto ao cálculo do critério de otimização, os trabalhos de maior destaque na comunidade utilizam a definição de entropia de Renyi [8] para propor um estimador universal de entropia que permite aplicar algoritmos de otimização dos parâmetros com busca pelo gradiente [1]. A entropia de Renyi pode ser vista como um caso geral da entropia de Shannon e até então mostra-se mais simples, com $\alpha = 2$, para derivação de um estimador eficiente computacionalmente. Todavia, além da eficiência e dos bons resultados empíricos, ainda não há argumentos teóricos que justifiquem a escolha da entropia de Renyi em detrimento da definição clássica de Shannon ou de outras definições alternativas.

Por isso também existem trabalhos que derivam estimadores baseados na entropia de Shannon e apresentam aplicações práticas [9, 13], embora ainda representem uma menor parcela dentro dos resultados práticos de ITL.

4.2. Caso discreto

A estimativa de uma função massa de probabilidade é consideravelmente mais simples que o análogo contínuo. Tendo um conjunto finito de N observações independentes e identicamente distribuídas, pode-se utilizar o seguinte estimador consistente:

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} \phi\{x_i = x\}, \tag{6}$$

onde ϕ é a função indicador e x_i uma observação. Com a estimativa das probabilidades estabelecida, calculam-se diretamente os valores de entropia ou informação mútua através das definições dadas nas Equações 1 e 3.

5. Considerações finais

O Aprendizado Baseado em Teoria da Informação é uma área de pesquisa bastante nova e que já apresenta resultados promissores, extrapolando o paradigma da otimização pelo erro quadrático médio ou por outras estatísticas de segunda ordem como a variância e correlação.

Problemas de natureza não-linear e com distribuição dos dados não obrigatoriamente gaussiana podem atualmente ser abordados por ITL de uma forma mais robusta. Mas por ser um campo de estudo novo, muitas questões ainda permanecem em aberto e assim fornecem oportunidades para contribuições:

- Não há um consenso sobre qual é a melhor abordagem para cálculo da entropia, se pela definição de Shannon ou se pela generalização de Renyi, o que dá oportunidade para se estudar comparativamente o desempenho de algoritmos adaptativos com os dois métodos. Além disso, precisa-se investigar mais profundamente o motivo da escolha da definição de entropia de Renyi.
- Os algoritmos de treinamento com os estimadores de entropia até agora possuem

complexidade $O(N^2)$ em função do número de amostras, enquanto que os algoritmos clássicos de treinamento (ex.: gradiente descendente) em batelada possuem complexidade $O(N)$. Logo há o desafio de aprimorar os estimadores de entropia para ganhar eficiência computacional com garantia de precisão.

- A aplicação de ITL necessita ser ampliada para outros problemas a fim de que se saiba se o seu uso é de fato superior frente a estatísticas de segunda ordem, seja em problemas já solucionados ou seja em problemas de maior complexidade e que ainda não possuem soluções satisfatórias pelos critérios tradicionais.
- Deve-se ampliar o o escopo de ITL para dados com valores discretos e soluções não restritas a algoritmos de aprendizado baseados em gradiente, o que possivelmente trará à tona abordagens inovadoras para problemas atualmente complexos ou até mesmo inéditos e que fogem do cenário de valores contínuos.

Este artigo, que havia sido apresentado em parte no Primeiro Simpósio de Processamento de Sinais da Unicamp [11], descreve de maneira bastante resumida esta nova área de pesquisa, sua definição e motivação. Para se aprofundar, são recomendadas as leituras dos trabalhos de Principe et al. [7] e Erdogmus [1, 3].

Referências

[1] D. Erdogmus. *Information Theoretic Learning: Renyi's Entropy And Its Applications To Adaptive System Training*. PhD thesis, University of Florida, 2002.

[2] D. Erdogmus and J.C. Principe. An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems. *IEEE Transactions on Signal Processing*, 50(7):1780 – 1786, 2002.

[3] D. Erdogmus and J.C. Principe. From linear adaptive filtering to nonlinear information processing. *IEEE Signal Processing Magazine*, 23:14–33, 2006.

[4] H.W. Gutch, P. Gruber, and F.J. Theis. Ica over finite fields. In *Proceedings of the 9th international conference on Latent variable analysis and signal separation*, pages 645–652. Springer-Verlag, 2010.

[5] S. Haykin, editor. *Unsupervised Adaptive Filtering: Blind Source Separation*. Wiley, 2000.

[6] T. Lehn-Schiøler, A. Hegde, D. Erdogmus, and J.C. Principe. Vector quantization using information theoretic concepts. *Natural Computing*, 4(1):39–51, 2005.

[7] J.C. Principe, D. Xu, and J. Fisher. *Information theoretic learning*, chapter 7, pages 265–319. Wiley, 2000.

[8] A. Renyi. *Probability Theory*. North-Holland, 1970.

[9] N.N. Schraudolph. Gradient-based manipulation of nonparametric entropy estimates. *IEEE Transactions on Neural Networks*, 15(4):828–837, 2004.

[10] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[11] D. G. Silva and R. Attux. Aprendizado baseado em teoria da informação: Fundamentos e perspectivas. In *Anais do Primeiro Simpósio em Processamento de Sinais da Unicamp*, 2010.

[12] V. Sindhwani, S. Rakshit, D. Deodhare, D. Erdogmus, J.C. Principe, and P. Niyogi. Feature selection in MLPs and SVMs based on maximum output information. *IEEE Transactions on Neural Networks*, 15(4):937–948, 2004.

[13] Paul Viola, Nicol N. Schraudolph, and Terrence J. Sejnowski. Empirical entropy manipulation for real-world problems. In *Neural Information Processing Systems 8*, pages 851–857. MIT Press, 1996.

[14] Z. Wang and A.C. Bovik. Mean squared error: love it or leave it?-a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009.

[15] A. Yeredor. Ica in boolean xor mixtures. *Independent Component Analysis and Signal Separation*, pages 827–835, 2007.

[16] A. Yeredor. Independent component analysis over galois fields. *Arxiv preprint arXiv:1007.2071*, 2010.