Ajuste de um modelo de campo médio para transdução semi-supervisionada

Emílio Bergamim Júnior
Fabricio Breve
IGCE/UNESP
Rio Claro, Brasil
emiliobergjr@gmail.com

Resumo*—Utilizando uma aproximação de campo médio para o modelo Potts, propomos dois procedimentos de ajuste para o parâmetro β de tal modelo, voltados para a tarefa de transdução semi-supervisionada. Visa-se estudar tal abordagem e compará-las com algoritmos do estado da arte em conjuntos de *benchmark* de forma a situá-la entre os demais métodos. *Área: Inteligência Computacional

I. Introdução

A transdução semi-supervisionada é uma sub-área do aprendizado semi-supervisionado focada no problema de classificar um conjunto de dados D dotado de um subconjunto D_l de instâncias cuja classificação correta é previamente conhecida [1]–[3].

Recentemente, modelos de campo médio oriundos da física estatística foram estudados em aplicações transdutivas, obtendo resultados comparáveis ao estado da arte [4]–[6]. Tais trabalhos chamam atenção para a acurácia desses modelos na tarefa em questão, ao mesmo tempo em que destacam a dificuldade de ajustar o parâmetro β de forma a extrair os melhores resultados possíveis.

Neste trabalho propomos duas maneiras de ajustar o parâmetro em questão delineando uma condição necessária para o aprendizado. A partir daí, estas três abordagens podem ser comparadas de forma a compreender seu desempenho na tarefa de transdução e, em seguida, comparadas com outras metodologias estabelecidas na literatura.

II. Conceitos e Técnicas

Seja $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ um conjunto no qual $\mathbf{x}_i \in \mathbb{R}^d$ e $y_i \in \{1, ..., q\}$ são respectivamente um vetor de atributos e o rótulo da $i-$ésima instância de D. Em um problema de transdução semi-supervisionada, existe um subconjunto $D_l \subset D$ para os quais os rótulos são previamente conhecidos e o objetivo é rotular todos os elementos de D [1], [3].

A primeira etapa do problema é então construir uma matriz de similaridade simétrica $\mathbf{W}_{N\times N}$ para a qual $W_{i,j}$ é a similaridade entre as $i-$ésima e $j-$ésima instâncias do conjunto.

Considere então variáveis aleatórias $\mathbf{s} = (s_1, ..., s_N)$ tais que $s_i \in \{1, ..., q\}$ representa o rótulo da $i-$ésima instância de D. A hipótese fundamental do aprendizado semi-supervisionado é que as diferentes classes presentes no conjunto estão relacionadas com uma estrutura de *clusters* no mesmo [1], [3].

Então, faz-se necessário um modelo probabilístico que seja capaz de capturar tal comportamento. Em particular, utilizaremos o modelo Potts [7]

$$p(\mathbf{s}) = \frac{1}{Z}\exp\{-H(\mathbf{s})\}, \tag{1}$$

sendo Z uma constante de normalização (também chamada função de partição) e

$$H(\mathbf{s}) = -\beta\sum_{i<j} W_{i,j}\delta(s_i, s_j) - \sum_{i,s}\theta_{i,s}\delta(s_i, s), \tag{2}$$

onde $\delta(\cdot,\cdot)$ é o delta de Kronecker e $\boldsymbol{\theta}_{N\times q}$ é uma matriz associada à informação *a priori* fornecida por D_l e dada por

$$\theta_{i,s} = \begin{cases} 1, \text{ se } (x_i, y_i) \in D_l \text{ e } y_i = s \\ 0, \text{ caso contrário.} \end{cases} \tag{3}$$

O termo dependente de $\mathbf{W}$ descreve a tendência de suavidade na qual pontos mais similares tendem a possuir o mesmo rótulo.

A. Naive Mean Fields

Modelos como o descrito em (1) são intratáveis devido à dificuldade em calcular Z. No caso do modelo Potts, por exemplo, isso envolve uma soma de complexidade $O(q^N)$, o que é inviável para N suficientemente grande [8].

Uma forma de contornar tal problema é através da utilização de aproximações de campo médio [8], [9]. Neste trabalho focamos em uma das abordagens mais simples, conhecida como *Naive Mean Fields* (NMF) [8], a qual aproxima (1) por uma distribuição $\Phi(\mathbf{s})$ que minimiza a divergência de Kullback-Leibler

$$D_{KL}(p||\Phi) = \sum_{\mathbf{s}} p(\mathbf{s}) \ln\left\{\frac{p(\mathbf{s})}{\Phi(\mathbf{s})}\right\}. \tag{4}$$

Em particular, adotando $\Phi(\mathbf{s})$ como sendo uma distribuição que é o produto de distribuições marginais

$$\Phi(\mathbf{s}) = \prod_i \phi_i(s_i), \tag{5}$$

X WORKSHOP DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO DA UNESP (WPPGCC 2021)

Bolsista CAPES

a minimização de (4) resulta em distribuições marginais que satisfazem um conjunto de equações não-lineares que podem ser resolvidas iterativamente. Ao final desse processo, as instâncias de D são classificadas de acordo com seu estado mais provável:

$$y_i = arg\max_{s_i} \phi_i(s_i). \tag{6}$$

B. Ajuste de β

Foi observado previamente que o valor de β escolhido para inferência decresce conforme aumenta-se o número de elementos em D_l [10]. Sendo $r_l = |D_l|/|D|$, definimos

$$\beta_i^* = \frac{1 - r_l}{\sum_{j \neq i} W_{i,j}} \tag{7}$$

e estudamos duas formas de ajuste, dadas por

$$\beta_{max}^* = \max_i \beta_i^* \text{ e} \tag{8}$$

$$\beta_S^* = \frac{1}{N} \sum_i \beta_i^*. \tag{9}$$

III. Metodologia de Desenvolvimento

De forma a avaliar os procedimentos de ajuste propostos, utilizamos três conjuntos de dados bidimensionais (duas luas, três clusters não-homogêneos e cinco gaussianas) e outros quatro conjuntos comumente utilizados para *benchmarks* de algoritmos de classificação (Digit1 [3], COIL [3], Mnist [11] e USPS [12]). A avaliação é feita em termos da acurácia das classificações e do tempo de execução dos algoritmos e como *baselines* são utilizados os algoritmos *Label Propagation* (LP) [13] e *Local and Global Consistency* (LGC) [14].

Para construção da matriz de similaridade $\mathbf{W}$ são utilizados os algoritmos PNN [15] e a similaridade gaussiana (RBF), como utilizada em [16]. Para construção do grafo de similaridade (isto é, determinação das entradas nulas de $\mathbf{W}$), é utilizado o método de k vizinhos mais próximos, sendo que tal parâmetro é variado em $k = 2, 3, ..., 20$.

Por fim, é relevante também avaliar o algoritmo sob diversas realizações de D_l. Para cada conjunto são então construídos vinte *splits* aleatórios contendo ao menos um representante de cada classe conforme varia-se o parâmetro r_l.

IV. Resultados Preliminares

Na Figura 1 são apresentados os resultados preliminares utilizando as similaridades PNN e RBF conforme o número de vizinhos mais próximos é variado. Nota-se que os resultados obtidos pela construção PNN são mais estáveis quando comparados aos de RBF, principalmente em valores menores de k. No entanto, o ajuste com β_S^* mostra-se promissor, visto que seu tempo de execução chega a ser uma ordem de magnitude menor que os demais algoritmos.

V. Considerações Finais

A sequência do trabalho visa executar os algoritmos discutidos nos demais conjuntos de *benchmark* para classificação, assim como compreender melhor o impacto das similaridades sobre os mesmos.

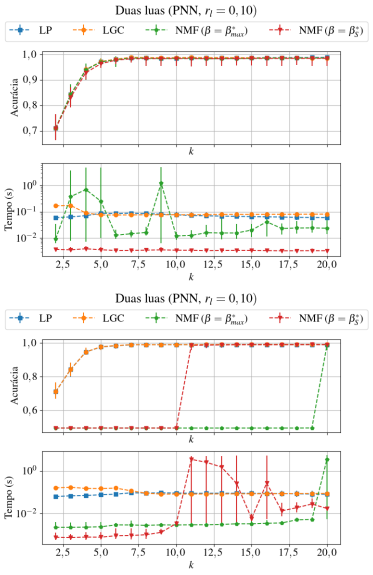

Figura 1. Resultados preliminares obtidos para o conjunto de duas luas.

Referências

[1] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.

[2] Y. Chong, Y. Ding, Q. Yan, and S. Pan, "Graph-based semi-supervised learning: A review," *Neurocomputing*, vol. 408, pp. 216–230, 2020.

[3] O. Chapelle, B. Scholkopf, and A. Zien, *Semi-supervised learning*. MIT Press, 2006.

[4] F. Wang, S. Wang, C. Zhang, and O. Winther, "Semi-supervised mean fields," in *Artificial Intelligence and Statistics*, 2007, pp. 596–603.

[5] J. Li and F. Wang, "Semi-supervised learning via mean field methods," *Neurocomputing*, vol. 177, pp. 385–393, 2016.

[6] C. Shi, Y. Liu, and P. Zhang, "Weighted community detection and data clustering using message passing," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2018, no. 3, p. 033405, 2018.

[7] F.-Y. Wu, "The potts model," *Reviews of modern physics*, vol. 54, no. 1, p. 235, 1982.

[8] M. Opper and D. Saad, *Advanced mean field methods: Theory and practice*. MIT press, 2001.

[9] J. S. Yedidia, W. Freeman, and Y. Weiss, "Generalized belief propagation," *Advances in neural information processing systems*, vol. 13, pp. 689–695, 2000.

[10] G. Getz, N. Shental, and E. Domany, "Semi-supervised learning–a statistical physics approach," in *In "Learning with Partially Classified Training Data"–ICML05 workshop*. Citeseer, 2005.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[12] J. J. Hull, "A database for handwritten text recognition research," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16, no. 5, pp. 550–554, 1994.

[13] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," 2002.

[14] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in neural information processing systems*, 2004, pp. 321–328.

[15] J. Ma, B. Xiao, and C. Deng, "Graph based semi-supervised classification with probabilistic nearest neighbors," *Pattern Recognition Letters*, vol. 133, pp. 94–101, 2020.

[16] C. A. R. de Sousa, S. O. Rezende, and G. E. Batista, "Influence of graph construction on semi-supervised learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2013, pp. 160–175.