攻读博士学位个人研究陈述书

Mincong Luo
China Institute of Atomic Energy
Peking,China
luomincentos@ciae.ac.cn

Abstract

我来自中国原子能科学研究院应用数学专业, 在硕士期间我主要从事智能算法研究. 不同于主流的人工智能算法, 我的工作更偏向数学而非计算机应用. 博士期间我打算继续在数学理论方面加强学习, 在研究工作上我准备从三方面入手, 争取在统计学和生物统计方面取得有意义的研究进展. 毕业后我仍然打算继续从事科研, 长远目标是组建自己的研究小组, 期望在统计智能算法的分支上做到一流;

全文组织结构

按照博士生报考材料的要求和我自己的叙述逻辑, 全文组织如下:

- **学术背景:** 从宏观角度介绍了我以往的研究方向和特点.
- **在所申请专业领域作过的研究工作:** 比较详细地介绍了我以往的研究内容及其意义和方法.
- **攻读博士生阶段的学习和研究计划:** 攻读博士期间我打算进行的研究方向.
- **以及博士生毕业后的就业目标:** 从短期和长远目标介绍了我之后的科研计划及职业生涯规划.

学术背景

我本科时的专业是核工程, 属于工程物理, 硕士时的专业是应用数学, 因此我比较喜欢从物理模型里抽象其数学本质来解决问题. 我一直致力做的是统计智能算法, 在研究中包括但不限于使用和统计方法相关、和生物原理相关的方法解决需要智能推理的问题.

0.1 统计智能算法

统计智能算法是一个大类, 运用**数学统计方法**对**生物智能行为**进行建模都属于统计智能算法, 如蚁群算法、PSO 粒子群算法、BP 神经网络等等 Stützle & Hoos (1999) Zhang et al. (2007) Hornik et al. (1990); 大多数统计智能算法可以表达为 $\mathbb{P}_\theta(y|X)$, 其中 θ 是我们想求取的模型参数,y 是我们的模型输出的预测,X 则是模型输入;

统计机器学习 Bishop (2006) 也属于统计智能算法, 其依赖数据集 X 进行有监督/无监督学习对参数 θ 做误差梯度下降, 但是也有很多自适应智能算法 (如 Hebb Rule 201 (2010)) 不依赖误差梯度下降; 因此当前的主流 AI 算法只是统计智能算法的真子集;

0.2 相关工作: 文献的智能挖掘推理

这是我硕士毕业论文的主要工作, 主要目的是用统计算法对海量文献的知识图谱进行建模, 流程: 数据集爬取自 arxiv.org 上的文章, 对文章文本分词, 并用类 word2vec 的算法计算其

特征向量; 这样对第 i 篇文章我们就可以求得其文中关键概念词的特征向量集合 S_i(相当于一个独立向量空间), 一个关键是将多个空间的特征向量映射到一个公共空间, 并保持其特征不变; 目前提出映射算法如下:

Algorithm 1 特征映射算法

输入: 来自多个论文文本的特征向量集的集合 $S = \{S_1, S_2...S_m\}$**, 其中** S_i **对应** $V_i = v_{i1}, v_{i2}...$

输出: 属于一个总体空间的特征向量集合 $V_{total} = \{v_1, v_2...v_s\}$

1: 对于容量最大的集合 $S_{max} \in S$;
2: **while** 可遍历概念集合 $S' \in S$ 并且 $S' \neq S_{max}$ **do**
3: if $s'_i \in S' = s_i^m \in S_{max}$: 使用映射 $v_k = f(v_i, v'_k, v'_i)$ 至特征空间并加入 V_{total}, 其中 $s_k \in S'$, 其中 v'_i 是概念 s_i 在自身特征空间的表达;
4: **end while**

0.3 相关工作: 基于物理数学的启发式算法

这是我研究生期间的研究工作, 基于统计物理学里的一些基础理论对智能算法做一些改进: 我的工作并非第一次做这个方向的尝试, 之前有基于统计物理学的学习算法 Boltzman Machine,Softmax 层等 Hopfield (1987) Mikolov et al. (2013); 我做过基于 Gibbs 过程的股票趋势分析 Luo (2018a), 和统计学习相关的有最大熵原理 (MEM) 的学习算法 Luo (2018b), 在 MNIST 数据集上取得了不错的效果; 简单来说, 我们通过证明如下定理:

Theorem 0.1 *信息熵的最大值为:*

$$S_{Imax} = \lambda_0 + \sum_r \lambda_r < f_r(x) >$$

然后结合神经网络的算法, 凑出了如下损失函数:

$$\begin{aligned} S_{Ip} &= -\sum_i p_i\{-\lambda_0 - \sum_r \lambda_r f_r(x_i)\} \\ &= \lambda_0 + \vec{\lambda} \odot (\vec{p} \otimes f(\vec{x})) \end{aligned}$$

0.4 关于我的研究背景的总结

总体来说我一直致力于研究统计智能算法, 在硕士期间取得了一些阶段性成果, 生物统计学的挖掘算法 (比如质谱鉴定蛋白质算法 Chi et al. (2018)) 是统计智能算法的一种特定应用场景, 我也希望之后能研究一些超越应用场景的通用算法.

在所申请专业领域作过的研究工作

我申请的是统计学的生物统计方向, 我计划用基于统计学的智能算法做生物统计学里的一些数据挖掘工作; 虽然之前没有做过专门针对生物信息学的智能算法研究, 但是核心的算法并不受制于应用场景和数据; 硕士期间我做过以下一些关于智能算法的研究工作:

0.5 基于注意力机制和先验知识的材料图像生成模型

生成模型相比统计判别模型, 区别在于是对真实分布 $\mathbb{P}(X)$ 采样以拟合一个分布 $\mathbb{P}(\hat{X}, z)$ 来模拟真实分布, 而判别模型是以 $\mathbb{P}(y|X)$ 在预测输入 X 的类别标签 y;

而材料图像生成模型则将先验知识先表达为特征向量, 然后再通过注意力机制 Vaswani et al. (2017), 使用 GAN Goodfellow et al. (2014) 生成材料辐照实验的图像, 构建 data-to-image 的生成模型是非常有意义的: 它使研究人员可以跳过危险和漫长的辐照实验根据材料特性 D_d 和实验工况 D_c 直接获得实验结果的图像 X_{img}.

0.6 正交策略梯度

相比统计生成模型和统计判别模型, 强化学习是去拟合学习一个 $\pi(a^{(t+1)}|s^{(t)})$ 的分布: 由状态空间的输入 $s^{(t)} \in S$ 求合适的策略空间的输出 $a^{(t+1)} \in A$;

首先由策略梯度的一个表达形式的引理出发, 即 $\frac{\partial \rho}{\partial \theta} = \sum_a d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a)$, 得出需要最小化的目标函数:

$$\frac{\partial \rho}{\partial \theta} \to 0 = \sum_a d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a)$$

这样可以用神经网络分别拟合 $\frac{\partial \pi(s,a)}{\partial \theta}$ 和 $Q^\pi(s,a)$ 并且使目标函数趋近 0(相当于两个向量正交), 这样构造出的策略梯度可以使得强化学习算法中的 Reward 函数值最大化.

0.7 模糊符号计算

符号计算其实是就是数学公式的自动推导以及数学定理的机器证明, 这个领域的研究一直是用逻辑规则推导在做, 而我一直尝试用统计智能算法在做符号计算, 这种基于软规则的计算可称之为模糊符号计算;

我提出 Graph Embedding 的方法来将公式表达为有向图 G, 再嵌入到特征向量空间 (Graph2Vec A & J (2016)), 然后用 Pattern Matching Network 来做公式匹配和推导, 最后用 $S_G(F'_r, F_r) = \lambda_0/(|G_s \in F_r \wedge G_s \notin F'_r| + |G_s \in F'_r \wedge G_s \notin F_r|)$ 来度量由模型推导出的结果 F'_r 和正确结果 F_r 的误差, 取得了不错的推导效果;

0.8 基于语义和视觉的逻辑推理

为认知建立数学模型对通用人工智能意义重大, 根据脑认知的原理可以将认知分为 3 个层次: 对具象的模式识别分类 $\mathbb{P}(y|X)$, 对抽象概念的学习 (Concept Learning) 包括概念识别 $\mathbb{P}_w(c|X)$ 和概念生成 $\mathbb{P}_w(X|c)$(其中 w 是具体概念的特征向量), 根据对经验的认知进行逻辑归纳 $\mathbb{P}(G_L|G_S^{(t)})$ 和推理 $\mathbb{P}(G_S^{(t+1)}|G_L)$(其中 G_L, G_S 分别是关于逻辑和关于认知的有向无环图); 我所做的主要工作主要是后两项;

关于 Concept Learning 我提出基于语义和视觉来学习概念的算法, 并证明了算法的梯度下降收敛性; 简要说交替训练概念识别网络 $p(y_a|X_i, X_q, A_s^{(t-1)})$ 和概念生成网络 $p(y_p|X_i, X_s, A_v^{(t-1)})$, 损失函数如下:

$$\mathcal{L}(A_s) = -\hat{y} \odot log(f_{softmax}(A_s \odot C^{*T})) + \lambda \frac{M||I||^2}{\nabla^2_{A_s} \mathcal{L}(\hat{\mathcal{A}}_f)}$$

我们证明的定理如下, 这个定理保证损失中的正则项会让参数迭代方向是更优化的;

Theorem 0.2 $\hat{\mathcal{L}}(A + \alpha\Delta A)$ 的一个上界的证明:$\hat{\mathcal{L}}(A + \alpha\Delta A) \leq \hat{\mathcal{L}}(A + \alpha\Delta A) + (\alpha - \alpha^2 M/2)||\nabla_A \hat{\mathcal{L}}||^2 \leq \hat{\mathcal{L}}(A)$

1 攻读博士生阶段的学习和研究计划

总体来说, 我希望在之前研究的一些阶段性成果的基础上, 夯实数学理论基础, 拓宽专业知识, 并在统计智能算法特别是针对生物统计学的统计智能算法上找到一个更细的方向, 兼顾数学理论深度和编程实现效果, 做出有意义的研究;

1.1 具有推理能力的统计智能算法

首先在硕士阶段我利用编程实现了一个名为 Fuzzy World 的虚拟 3D 工具, 蕴含了一个封闭世界的逻辑规律, 可以检验智能算法是具备真正的复杂逻辑推理能力; 我将世界的状态和蕴含的逻辑抽象为语义图 (有向无环图):

- **环境状态认知语义图** G_S**:** 谓词连接了实体或概念, 所以 G_S 是多个三元组组成的, 比如 $is(moving, cow)$ 或者 $biggerThan(cow, desk)$.
- **逻辑推断语义图** G_I**:** 是有因果联系的事件构成的, 比如 $is(openning, light) \Rightarrow is(hot, room)$ 就是一个一阶逻辑的事件关联.

基于强化原理和深度神经网络我们可以实现如下模块, 最终组合这些模块, 就可以构造具有完备推理能力的 human-level Agent:

- 用于决策的 policy network $\pi(a^{(t)}|G_S, G_I, l^{(t)})$;
- 用于环境物体状态识别的 recognization network $R(v^{(t)}, l^{(t)})$;
- 用于构建逻辑关系的 inference network $I(G_S^{(t)}, G_S^{(t-1)}, v^{(t)})$;
- 用于评估策略的 value network $Q(A, G_S, G_I, l^{(t)})$;

对于构造具有推理能力的统计智能算法我目前已有如上框架, 但是仍然需要更多统计和凸优化数学理论来使得算法的合理性及收敛性得到优化, 希望能在博士期间从事这一有意义的研究;

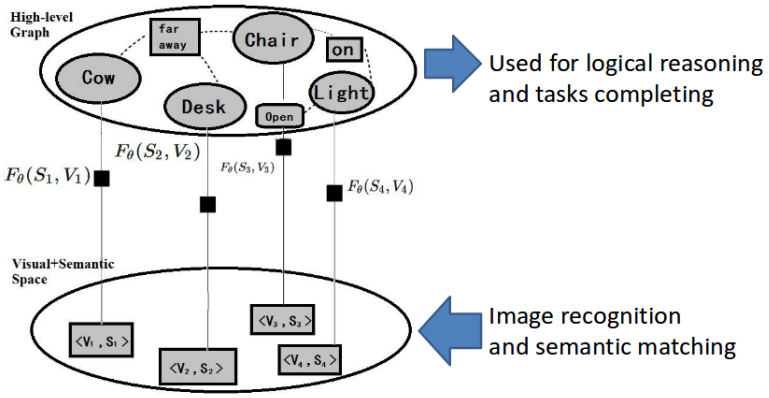

Figure 1: 这是统计算法和符号逻辑的结合, 统计算法可以将底层信号转化为图结构 G_S, 然后再进行符号逻辑运算.

1.2 用于生物信息学的智能算法

生物信息学中很多问题可以很容易地建立为智能算法模型, 比如蛋白质质谱鉴定问题 (质谱图相当于直方图, 目标是从这个数字化质谱图反推生成它的蛋白质) Chi et al. (2018), 可以将质谱 $X \in \mathbb{R}^d$ 进行特征提取为 $\hat{X}$(比如用简单的 SVM 提取 $X = u\hat{X}d^T$), 然后再构造判别模型 $y = f_\theta(\hat{X})$, 关于判别函数的构造我有一个之前做放射性能谱识别研究的思路 (和蛋白质质谱鉴定没有本质区别, 都属于模式识别类): 使用弱分类器集合提高模型的准确率和泛化性, 即 $\hat{y} = \sum_i \widetilde{w}_i f\theta_i(\hat{X})$;

除了上述这种在计算级别的算法改进, 还有更智能的算法可以革命性地改变生物信息学的研究, 我们可以构建一个从自然语言到知识图谱的学习器 $F(G|X_L)$, 这个基于统计自然语言的模型可以学习大量的生物学文本文献集合 X_L, 最终可以给出知识图谱 G, 利用 G 可以进行生物实验信息挖掘, 生物文本信息挖掘, 计算生物中算符和文本概念实体联系的挖掘等等, 简单来说我们要做的相当于类脑智能的生物专家知识系统; 事实上这是我硕士毕业论文的工作, 目前我在做的是原子核物理里的知识实体抽提算法, 但是这是一个和应用领域无关的算法, 这是一个在自然语言 (语言符号 encoding) 和特征知识 (特征向量空间) 之间建立映射的研究;

在这个小节里我列举了不同细粒度的两种可推进生物信息学研究的智能算法 (**针对特定底层算法级别**和**具有泛化性的专家系统**), 我相信这都是很有意义的研究工作.

1.3 一些偏应用工程的工作

除了理论研究之外, 我希望完成一些偏计算机工程的工作, 意义是减少重复性的工作, 提高科研的效率; 比如在硕士期间我开发了一些可重用的工具和数据集, 包括用于实验强化学习算法的 Fuzzy World, 以及数据集 ISMD Luo et al. (2018) 和 DESD Luo & Liu (2018).

提升科研效率的核心之一就是将实验流程自动化, 当想到一个算法时, 我们先开始做一些理论证明和推算工作 (这一步可以由自动符号计算替代一部分工作), 然后将这套算法在数据集

上进行实验, 事实上第二步的重复性很高. 可以建立一个 Linux 工作站, 上面可以内置常用的计算库和数据集, 并且每个拥有权限的人都可以提交自己的算法, 将算法和实验解耦, 可以大大减少重复性工作使得不必每次都从头开始. 编程语言上也应该统一, 应该限制在 Python/R 等语言, 使得至少研究小组之间的工作可以共享;

计算机工程的优点在于自动化一些可重复的工作, 在博士期间我希望发挥一些之前的编程经验优势提升科研效率.

2 博士生毕业后的就业目标

2.1 短期目标: 从事一到两个博士后研究

在攻读博士时, 我期望我能在理论上钻研得足够深入, 选择博士后研究时我期望能做一些偏 Application 的工作. 申请世界顶级科研小组的博士后是我之后的目标, 方向上我青睐将统计智能算法用于实业 (硬件, 制造业,Robotics 方向) 的研究;

2.2 长远目标: 组建自己的科研团队

组建自己的科研团队并从事某一个更细致方向的研究是我的长远目标, 我希望能在顶级科研机构立足并逐渐领导一支对科研有激情, 成员精通方向多, 科研产出质量高, 并具备培养研究生能力的领域内同行高度认可的团队.

3 总结

在这篇个人陈述中我从学术背景, 学术经历及博士计划等方面阐述了对研究工作的认知和定位, 总的来说我会结合数学理论和应用背景做出有意义的研究工作.

References

Hebb Rule. Springer US, 2010.

Grover A and Leskovec J. node2vec: Scalable feature learning for networks. In *Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 855–864, 2016.

Christopher M Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.

Hao Chi, Chao Liu, Hao Yang, Wen-Feng Zeng, Long Wu, Wen-Jing Zhou, Xiu-Nan Niu, Yue-He Ding, Yao Zhang, Rui-Min Wang, Zhao-Wei Wang, Zhen-Lin Chen, Rui-Xiang Sun, Tao Liu, Guang-Ming Tan, Meng-Qiu Dong, Ping Xu, Pei-Heng Zhang, and Si-Min He. Open-pfind enables precise, comprehensive and rapid peptide identification in shotgun proteomics. *bioRxiv*, 2018. doi: 10.1101/285395.

Ian J. Goodfellow, Jean Pougetabadie, Mehdi Mirza, Bing Xu, David Wardefarley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3:2672–2680, 2014.

Hopfield. Learning algorithms and probability distributions in feed-forward and feed-back networks. *Proceedings of the National Academy of Sciences of the United States of America*, 84(23):8429–8433, 1987.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, 3(5):551–560, 1990.

M. Luo, Xinfu He, and Li Liu. Generative model for material experiments based on prior knowledge and attention mechanism. *NeurIPS2018 Workshop on MLMM*, abs/1811.07982, 2018.

Min Zhong Luo and Li Liu. Automatic derivation of formulas using reforcement learning. 2018.

Mincong Luo. Gibbs method for stock market trend prediction. 2018a.

Mincong Luo. Mem-nn: A classification nueral network based on maximum entropy method. 2018b.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *International Conference on Neural Information Processing Systems*, pp. 3111–3119, 2013.

Thomas Stützle and Holger Hoos. *The Max-Min ANT System and Local Search for Combinatorial Optimization Problems*. Springer US, 1999.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.

Jing Ru Zhang, Jun Zhang, Tat Ming Lok, and Michael R. Lyu. A hybrid particle swarm optimization–back-propagation algorithm for feedforward neural network training. *Applied Mathematics and Computation*, 185(2):1026–1037, 2007.