## 石化生产过程基于大数据解析的分类建模方法研究

1. **研究背景及意义**
   1. **软测量soft sensor**

软测量在工业数据中具有重要意义。工业数据中存在很多难以测量的过程量，如果利用易于测量的过程量对其进行建模的话，可以大大缓解其中测量成本。这样一方面可以减少传感器数量，另一方面保证减少了冗余数据，提高了计算性能。

软测量实际上是一种数学模型，不同于传统的物理传感器，软测量传感器是一种虚拟的传感器，是一种通过计算得到数据的传感器[1]。

* 1. **软测量方法soft sensor**

鉴于软测量的便利性和可行性，已经有很多工作建立在软测量工作上。主要的方法有基于机制的方法，基于知识的方法，基于数据驱动的方法[2]。在计算资源如此丰富的现代社会，面对复杂多样的工业数据，数据驱动方法逐渐占据了它所具有的优势。在上世纪90年代兴起的传统机器学习方法，现在已经应用在软测量的各个方面。PLS（偏相关回归）可以用来处理高维且高度相关的数据。但有研究表明传统的PLS可以会导致大m小n问题，影响模型的预测效果[3]。SVM作为一个广为周知的机器学习方法，在分类方面有很高的地位。有文献[4]指出SVM在石油催化裂化分馏塔数据中能起到的重要作用。但SVM的计算量和数据成平方关系[5]，这就意味着如果数据量很大的话，SVM的计算效率会很低。ANN（人工神经网络）面对复杂多样的数据，在训练时可以自动得到权重，避免了SVM和PLS的诸多不便[6]。近几年来，各式各样的神经网络层出不穷。深度学习方法逐渐展现他在软测量问题上的优势。一方面，深度学习方法可以综合考虑各种变量，不仅防止了很多无用变量的冗余，并且保证了每个变量的高效利用。同时，在GPU的加速下，深度学习方法的计算速度得到保证，实际应用场景广泛。总而言之，深度学习方法综合了各种机器学习的优点，一定程度上避免了它们的一些缺点。同时，深度学习方法可以根据不同的数据集设定不同的参数，可以应用于不同场合。

CNN可以被证明是用来提取特征的有效方法[7]，对于一维数据，CNN可以帮助理解局部片段信息。对于二维数据，CNN更能体现在局部特征提取的独特优势。相较于MLP的较大的计算量，CNN不仅节约了计算量，并且在加深网络上有很大的作用。Resnet[8]作为CNN的集大成者，在图像处理领域充当了很久的特征提取作用。类似的，在处理时序信息上，循环神经网络的应用越多来多。LSTM长短时记忆网络[9]在循环神经网络上，一定程度上避免了梯度爆炸和梯度消失，逐渐作为最广为人知的循环神经网络。

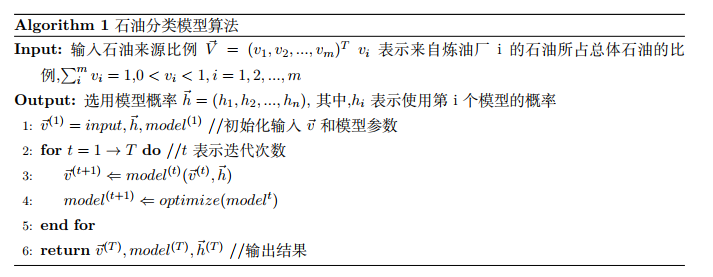
2017年横空出世的transformer[10]，利用它强大的自注意力机制，一举成为最具影响力的大模型，其应用广泛，在各个领域都大放异彩。目前，在NLP领域，有BERT等大模型[11]。在CV领域，有visual-transformer[12] 等突破传统CNN的应用。但是这些方法往往是具有普适性的，是基于方法和规则的，而不是基于数据类型的。面对工业过程，其数据的多样性让模型的鲁棒性备受考验。一般来说，基于数据集选择合适的模型和进行相应的参数调节是主要的流程方法。所以笔者想从实际工业数据入手，尝试将已有模型应用到工业生产中，在此基础上，发掘模型的鲁棒性，争取在面对更多类型的数据时让模型依然具有相当不错的效果。

* 1. **深度学习与软测量结合的挑战**

基于上述方法论述和模型应用场景及区别，总结一下深度学习方法在面对软测量的主要困难和挑战：[13]

1. 如何正确的应用深度学习的前沿模型，如何保证前沿模型的有效性能够落实到现在工业数据上？
2. 实际工业数据有很多形式的数据，如何根据数据类型匹配相应的模型，这是一个挑战。
3. 数据的噪声应该如何处理，怎么才能验证模型的有效性？
   1. **石油分类模型的重要性**

对于石油来说，往往一个比较好的分类模型是很有必要的[15]。对于石油数据来说，来自不同炼油厂的石油的材质有很大区别，如何根据不同炼油厂的石油的材质的区别选择合适的模型，是一个值得研究的任务。这就涉及到一个分类模型的问题，根据比例对石油分类，确定一个石油的分类模型。



具体操作流程如下，图片节选自[15]

|  |
| --- |
|  |
| 图1 : 石油分类模型的示意图 |

首先使用分类模型进行一步建模，得到石油的类别，然后根据石油的类别选择不同的模型进行下一步的建模操作。

* 1. **石油分类模型的挑战和问题**。

1.石油分类模型在实际应用过程中，可能没有这么多的数据，对于数据驱动的模型来说是一个很大的挑战。目前也有这种文献[16]提到如何处理这种数据不足的方法。

2.石油分类模型需要的特征数量少，输出信息不丰富，如果用很深的网络可能会有很严重的过拟合风险。如何使用合适的网络加强神经网络的非线性，保证拟合的准确性，是一个值得思考的问题。[17]提出了分段线性达到非线性的方法，很有借鉴意义。

3.如何确保神经网络在分类模型中的非线性问题，如何将实际工业数据能合理的应用到现有模型上是个比较困难重重的问题。

1. **目前的基于深度神经网络的研究进展**
   1. **基于LSTM和attention的软测量模型[14]**

LSTM模型可以注重时序逻辑，但不能关注不同时间的不同变量。注意力机制可以增进时间空间的联系，不仅加强了采集变量之间的关系，并且加强了变量和预测的质量变量的关系。模型结构如下图所示：

|  |
| --- |
|  |
| 图2： LSTM+attention的软测量模型 |

这篇文章发表是2021年发表的，力求通过attention的引入帮助解决单纯LSTM的一些缺点，训练资源要求不高，实验结果也要优于原来的一些方法，如DBN，ARIMAX，RNN等。但这些方法要求选择数据的速率是恒定的，也就是说，这种网络针对于时序信息更有优势。

此数据集是针对于粒度变化规律的，不是针对石油数据的，相较于石油数据，其时序逻辑会更贴切LSTM。

* 1. **基于深度信念网络(DBN)的分类模型**

Gao, Xiaoyong, et al[15]提到了用DBN方法，利用深度信念网络的非线性实现一种非线性映射，DBN的结构如下图所示，由很多个RBM构成（受限玻尔兹曼机）：

|  |
| --- |
|  |
| 图3： DBN的主要结构 |

在DBN中，主要是通过石油的来源比例来得到选用某个分类模型的概率，实际计算过程如下：

首先对于binary-binary 的RBM，有以下联合概率成立：

其中，都是可学习的参数，其中对应的条件分布为：

根据上述联合概率和条件概率，可以得到相应参数更新公式：

根据上述参数更新公式，可以得到每一层RBM的输出结果,然后把上一层的输出结果作为下一层RBM的输入。然后使用一种贪婪的训练方法——DBM的训练是从底部往上训练的。假设已经训练好了g层RBM，然后可以加上一层RBM，用训练好的g层的输出结果作为第g+1层的输入进行对第g+1层RBM的训练。这样的训练方式可以节约很多计算资源。

以上的训练方法都是在无监督的情况下训练的，因为可以看到这里的参数更新是通过最大化来实现的，没有使用到任何监督信息，这样可以增加模型的学习空间，一定程度上防止了过拟合。

无监督的模型训练好了之后，可以根据监督信息，即实际用到的模型编号来进行监督训练。假如原料比为的原油数据，它所用到的石油模型编号为j，那么监督信号是一个one-hot的向量，即，其他位置都是0。

假设DBM有L层RBM，那么训练得到的选用模型i的概率为：

也就是说，训练得到是使用模型i的概率是，如果在测试的时候的情况下，可以的方法来进行选择。同时，对于损失函数来说，由于是分类问题，可以用交叉熵损失函数。即：

这样就可以利用监督信息进行监督训练，减小损失，加强拟合能力。

* 1. **深度神经网络与石油数据的联系**

目前的深度神经网络样式极多，对于不同类型的数据，都能找到一些适配的神经网络。除了上述讲的两篇文章的方法，还有很多文章对于如何让深度学习真正落地到实际数据上有很深入的研究。例如：

1. 对于连续数据，按频率采样的数据，可以把输入数据认为是一个2D的矩阵，对于这种数据，可以认为是另一种形式的图片。那么，一些对于图片起作用的神经网络就可以应用上去。[14][18][19]等文章都从不同方面阐释了时序信息在软测量建模中的应用。使用到的数据集也都是一些工业数据集，可以很容易的推广到石油数据集上面
2. 神经网络的训练往往需要大量的计算资源，如果使用特别大的模型例如transformer[10]的话，可能会对于计算资源有比较大的负担。倘若没有GPU等计算能力强大的资源的话，普通的CPU计算资源容易导致计算速度过慢，难以应用到实际情况下。为此，简化网络和减少计算量也是一直存在的一个问题。[20][21]等文章分析了目前比较流行的软测量方法和现有炼油过程。为实际应用神经网络到石油数据上提供了一个良好的参考。
3. **研究内容**

基于上述研究背景和已有研究成果，我的研究内容有以下三点。

* 1. **熟悉仿真软件的使用和熟练掌握用仿真软件生成仿真数据**

HYSYS是世界著名油气加工模拟软件工程公司开发的大型专家系统软件。该软件分动态和稳态两大部分。其动态和稳态主要用于油田地面工程建设设计和[石油](https://baike.baidu.com/item/%E7%9F%B3%E6%B2%B9?fromModule=lemma_inlink)石化[炼油](https://baike.baidu.com/item/%E7%82%BC%E6%B2%B9/3198864?fromModule=lemma_inlink)工程设计计算分析。其动态部分可用于指挥原油生产和储运系统的运行。

《Guide to the Fluidized Catalytic Cracking Unit Standard Model》这本书详细介绍了如何用目前实验室的软件进行仿真数据生成的具体操作。可视化模型类似下图：

|  |
| --- |
|  |
| 图4 :可视化仿真软件架构 |

仿真软件可以生成足够多的数据用于数据驱动的模型训练，极大的减轻了因数据不足而产生的困难。生成仿真数据也是模型训练进行训练的第一步。

还有一些仿真软件没有可视化界面，要使用脚本语言进行操作。为此，要熟练掌握石油工业流程和石油原料等过程参数和过程装置。

* 1. **利用石油数据建立分类模型提高模型精度**

分类模型是把数据进行分类的模型，对于炼油数据，分类模型是针对原油数据进行分类，选择合适的模型进行使用[15]。由于数据特征数量较少，用很复杂的模型去拟合会导致过拟合现象的出现。为了保证模型的效果，我计划按照以下的流程进行研究：

1. 首先针对已有模型进行复现，[15][17]提到了用DBM和分段线性插值等方法进行石油分类模型的研究。其中，DBM的准确率很高，误差也比较小，如下图所示：

|  |
| --- |
|  |
| 图5： DBM和NN的误差比较 |

对于DBM和NN的结果差异，我认为主要是因为DBM对于非线性的拟合情况更好导致的。传统的NN可能在欠缺一些非线性的手段，非线性程度不高。但DBM是有良好的非线性的，这会在非线性的分类中产生更好的效果。如果在合理增加NN的深度的情况下，可以逐渐加强非线性的程度，达到类似DBM或者超过DBM的效果。所以可以在使用一些更深的网络进行拓展和改进现有模型。

1. 优化计算时间。对于神经网络来说，计算成本大。梯度下降和反向传播设计一些矩阵计算，很容易让时间成本过大。对于已有的模型，需要合理选择网络参数，在保证计算速度的情况下，保证网络的性能。
2. 在已有模型基础上，尝试用比较新的模型[8][10][12]进行改进，争取实现和已有的仿真软件合理接轨。
   1. **石油数据分类建模方向研究**

实际上面对的石油数据可能比较复杂，与单纯的石油组分相比，不仅是数据来源的复杂，更是数据特征维度的复杂。

一方面，很多基于大数据分析的工业处理的工作都是基于特定的数据集的，例如[22]是针对磨矿——溢流浆数据，但是[14]是面对加氢裂化石油过程中的数据的。尽管他们使用的模型都有LSTM[9]，但是确实面对不同的数据集。数据集的差异会导致模型的其他参数有很大不同，必须根据不同的数据集，相应的修改类似的模型。

另一方面，数据的来源面临着诸多困难。现有的数据集的大小往往难以支撑数据驱动的算法。[15][23]所用到的数据集都是不到10000的，这个数量级对于很多复杂的模型来说是稀缺的，当数据特征比较多是，稀缺的数据条数会让模型难以学习到更多的特征，陷入过拟合的境地。为此，必须用仿真软件设计更多的数据，让模型学习的层数更深，训练得到的效果更好。

这部分的研究重点在于：

1. 用已有模型（或者改进之后的模型）验证炼油数据分类建模过程中的可行性和有效性。包括训练和测试的准确率，训练时间花费等。
2. 结合仿真软件和实际工业数据，应用近5年来比较适配的神经网络，与之前的分类方法进行对比，选择适配的分类建模方法，在原有的分类建模方法上实现石化生产过程的模型特异性。
3. **工作计划**
4. 秋季学期12-17周，阅读软测量建模和石油分类模型文献，阅读仿真软件指导书，积极与老师沟通，学习仿真软件的使用，自己能够生成相应的仿真数据。
5. 春季学期1-3周，阅读相关文献，了解石油分类模型在提高模型精度方面的重要性，研究数据分类模型的基本方法。准备中期报告。
6. 春季学期4-8周，阅读相关文献。根据中期报告的结果，改进自己的工作，进行稳态建模。
7. 春季学期9-12周，阅读相关文献，进一步改进模型的细节，实现模型与仿真软件的匹配和耦合。
8. 春季学期13-16周，撰写毕业论文。
9. **参考文献**
10. Jiang, Yuchen, et al. "A review on soft sensors for monitoring, control, and optimization of industrial processes." *IEEE Sensors Journal* 21.11 (2020): 12868-12881
11. Kavuri, Rengaswamy Surya N . "A review of process fault detection and diagnosis: Part II: Qualitative models and search strategies." Computers & Chemical Engineering (2003).
12. Jiang, Qingchao , et al. "Data-Driven Batch-End Quality Modeling and Monitoring Based on Optimized Sparse Partial Least Squares." IEEE Transactions on Industrial Electronics PP.99(2019):1-1.
13. Yan, Weiwu, Huihe Shao, and Xiaofan Wang. "Soft sensing modeling based on support vector machine and Bayesian model selection." *Computers & chemical engineering* 28.8 (2004): 1489-1498.
14. Burges, Christopher JC. "A tutorial on support vector machines for pattern recognition." *Data mining and knowledge discovery* 2.2 (1998): 121-167.
15. Gonzaga, J. C. B., et al. "ANN-based soft-sensor for real-time process monitoring and control of an industrial polymerization process." *Computers & chemical engineering* 33.1 (2009): 43-49.
16. Horn, Z. C., et al. "Performance of convolutional neural networks for feature extraction in froth flotation sensing." *IFAC-PapersOnLine* 50.2 (2017): 13-18.
17. He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
18. Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.
19. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
20. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
21. Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).
22. Sun, Qingqiang, and Zhiqiang Ge. "A survey on deep learning for data-driven soft sensors." *IEEE Transactions on Industrial Informatics* 17.9 (2021): 5853-5866.
23. Yuan X, Li L, Shardt Y A W, et al. Deep learning with spatiotemporal attention-based LSTM for industrial soft sensor model development[J]. IEEE Transactions on Industrial Electronics, 2020, 68(5): 4404-4414
24. Gao, Xiaoyong, et al. "Refinery scheduling with varying crude: A deep belief network classification and multimodel approach." *AIChE Journal* 60.7 (2014): 2525-2532.
25. Wang, Ping, et al. "Adaptive nonlinear model predictive control using an on-line support vector regression updating strategy." *Chinese Journal of Chemical Engineering* 22.7 (2014): 774-781.
26. Gao, Xiaoyong, et al. "Optimizing scheduling of refinery operations based on piecewise linear models." *Computers & Chemical Engineering* 75 (2015): 105-119.
27. Geng Z, Chen Z, Meng Q, et al. Novel transformer based on gated convolutional neural network for dynamic soft sensor modeling of industrial processes[J]. IEEE Transactions on Industrial Informatics, 2021, 18(3): 1521-1529.
28. Yuan X, Qi S, Wang Y, et al. A dynamic CNN for nonlinear dynamic feature learning in soft sensor modeling of industrial process data[J]. Control Engineering Practice, 2020, 104: 104614.
29. Shang C, You F. Data analytics and machine learning for smart process manufacturing: recent advances and perspectives in the big data era[J]. Engineering, 2019, 5(6): 1010-1016.
30. 黄德先, 江永亨, 金以慧. 炼油工业过程控制的研究现状, 问题与展望[J]. 自动化学报, 2017, 43(6): 902-916.
31. Zhou J, Wang X, Yang C, et al. A novel soft sensor modeling approach based on difference-LSTM for complex industrial process[J]. IEEE Transactions on Industrial Informatics, 2021, 18(5): 2955-2964..
32. Shang C, Yang F, Huang D, et al. Data-driven soft sensor development based on deep learning technique[J]. Journal of Process Control, 2014, 24(3): 223-233.