基于海量专利数据对产业发展进行预测的算法对比

杨一扬（yyyang0817@163.com），左任衔（renxian@whu.edu.cn），

欧阳图(tu.ouyang@case.edu)

一、介绍

当前，国际科技与产业竞争尤为激烈。为了赢得未来全球产业发展先机、抢占世界科技竞争制高点，世界主要国家纷纷着眼未来，加强未来产业的前瞻布局。为预测在 2035 年左右可能发展成为未来产业的产业，我们采集了从 2000 年至 2021 年 8495.6 万条德温特专利数据，通过隐含狄利克雷分布（LDA）主题模型对专利的摘要信息进行聚类，聚类目的是为了能够筛选出较为全面反映全球未来技术研发趋势的技术族群，这样可使得预测出的未来产业前沿技术方向能够有更大可能性成长为战略性新兴产业乃至战略性支柱产业。聚类将德温特专利数据按中国、美国、全球分别聚类至困惑度指标最优的 20 个主题，以得到不同国家 20 个主要产业在 2000 年至 2021 年的发展情况。进而通过 Fisher-Pry 模型根据聚类结果对主要产业从 2021 年至 2050 年的发展趋势进行预测，以确定 2035 年左右可能发展成为未来产业的产业。因此在上述过程中，Fisher-Pry 模型对于数据的拟合精度决定了我们对于 2035 年产业预测的准确性，我们分别采用最小二乘法和模拟退火算法对 Fisher-Pry 模型进行求解，我们的实验结果显示用模拟退火算法求解达到的预测效果更好。

二、不同算法的 Fisher-Pry 模型预测实验

2.1 Fisher-Pry 模型介绍

Fisher-Pry 模型最早来源于 Fisher J.C.和 Pry R.H.于 1971 年发表的一篇合作论文，是目前评估技术成熟度相对准确率较高的方法之一，目前已有不少研究通过 Fisher-Pry 模型利用 SCI、EI、专利与商业报道等各种数据对技术发展进行预测，这进一步启发我们基于 Fisher-Pry 模型预测未来产业。具体地，Fisher-Pry 模型形式如下：

$$f(t) = \frac{K}{1 + e^{-b(t-a)}}$$

其中K代表最大专利数，t代表时间，a代表专利发展到最大专利数一半时对应的时刻，b代表专利发展速率。

2.2 基于最小二乘法的模型求解

最小二乘法是从线性拟合的角度求解 Fisher-pry 模型。首先对$f(t)$进行对数变化得：

$$-b(t-a)=ln\left[\frac{K-f(t)}{f(t)}\right],$$

等式左边是一个与t有关的线性函数，右边是与$f(t)$有关的对数函数，其中a，b是代拟合参数，K是已知参数。因此首先需要对K值进行估计，这里选择拐点法估计K值。即计算已知专利数据的梯度，取最大梯度处对应专利数的两倍作为K值：

$$K=2\times max\{\frac{f(t_{n+1})-f(t_n)}{t_{n+1}-t_n}\}\qquad n=1,2,3,\dots,N-1,$$

其中N是已知数据的截止时间，那么已知N对$\big(t,f(t)\big)$，令$C=ln\left[\frac{K-f(t)}{f(t)}\right]$，我们拟合如下曲线：

$$-b(t-a)=C,$$

即可得到拟合参数a，b。将上文估计的K值，拟合结果a，b代入到拟合目标中，即可得到求解后的 Fisher-Pry 模型。

2.3 基于模拟退火算法的模型求解

基于模拟退火算法的模型求解思路是利用模拟退火算法估计出最符合真实值的K，a，b，进而直接求解 Fisher-pry 模型。模拟退火算法是一种基于蒙特卡洛思想的非精确搜索算法，通过不断变化自变量的值，找到最优的函数结果。具体地，我们首先需要对代估参数K，a，b的范围做约束：对于K，a来说：

$$K\in\left[\frac{1}{2}\times max[f(t)],4\times max[f(t)]\right],\ a\in[2\times\min[t],2\times max[t]-min[t]],$$

对于b来说：

$$b=\frac{\ln[81]}{\Delta t},\ b\in\left[\frac{ln[81]}{max\{t\}-min\{t\}},\frac{8\times ln[81]}{max\{t\}-min\{t\}}\right]。$$

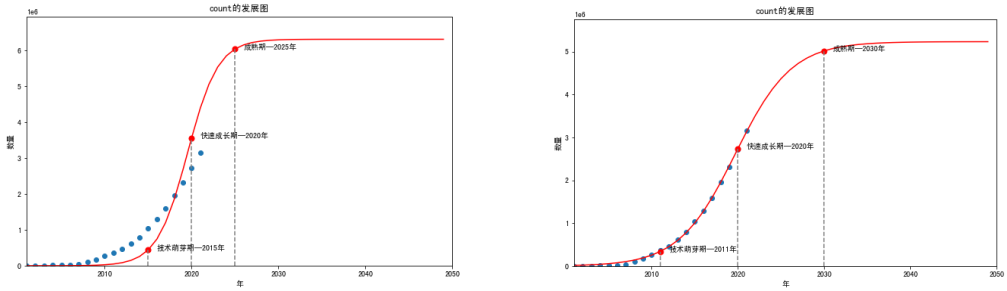
模拟退火的目标函数为:

$$goal = \sqrt{\left(f(y^*) - f(\hat{y})\right)}$$

其中$f(\hat{y})$为从约束范围内随机选定的一组K，a，b，代入至拟合目标中得到的函数结果，$f(y^*)$为真实的结果。因此目标函数$goal$越小，则选取的参数K，a，b越接近真实值，在退火过程中不断更新最小的目标函数值对应的参数，最终得到最符合真实结果的K，a，b，将其代入拟合目标中，即可求解后的 Fisher-Pry 模型。

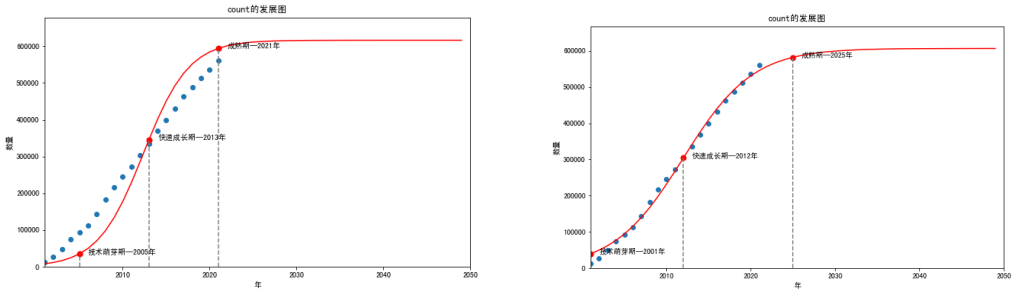
2.4 两种算法的对比结果：

本研究使用 LDA 主题模型，对专利的摘要信息进行聚类，生成 20 个主题，每个主题由 10 个单词构成。每个主题可以表现一个产业类型，同时每一条专利会在每一个主题上存在权重分布，因此每一个主题在每一年专利上的平均权重即可作为该产业在当年的发展程度，主题对应的德温特专利数量加权和可作为我们模型的待拟合数据。因此，我们分别利用基于最小二乘法的 Fisher-Pry 模型和基于模拟退火算法的 Fisher-Pry 模型对 LDA 主题模型得到的不同地区（即美国、中国和全球）某主要产业（即 20 个聚类产业之一）数据进行拟合，见图 1，2，3 所示。



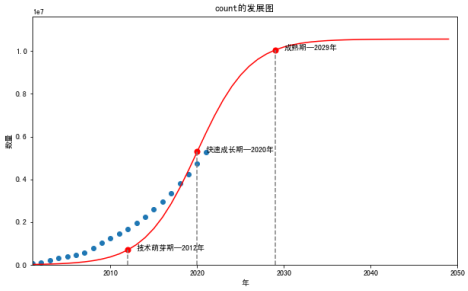
(a)最小二乘法　　(b)模拟退火法

图 1 中国某主要产业拟合数据。其中横坐标为年份，纵坐标为专利数量，蓝点表示该产业真实专利加权数量，红色为算法对该产业专利数量的拟合曲线。

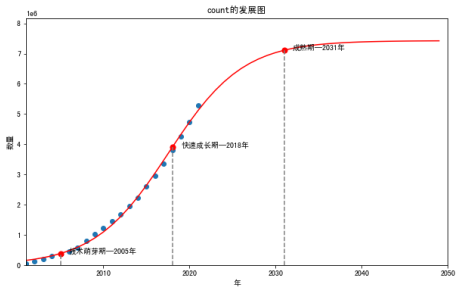


(a)最小二乘法　　(b)模拟退火法

图 2 美国某主要产业拟合数据。其中横坐标为年份，纵坐标为专利数量，蓝点表示该产业真实专利加权数量，红色为算法对该产业专利数量的拟合曲线。



(a)最小二乘法



(b)模拟退火法

图 3 全球某主要产业拟合数据。其中横坐标为年份，纵坐标为专利数量，蓝点表示该产业真实专利加权数量，红色为算法对该产业专利数量的拟合曲线。

我们分别用 MAPE(Mean Absolute Percentage Error)和确定系数 R2 度量拟合结果，其中 MAPE 越接近 0，R2 越接近 1，则拟合越接近真实情况。经过计算模拟退火算法的 MAPE 为 **0.16**，R2 为 **0.99**；最小二乘法的 MAPE 为 **0.51**，R2 为 **0.84**。为进一步对两种算法进行对比，我们随机选取 LDA 主题模型聚类对全球专利数据聚类之后的十个主题分别绘制两种算法的 MAPE 曲线和 R2 曲线，见图 4，5 所示。

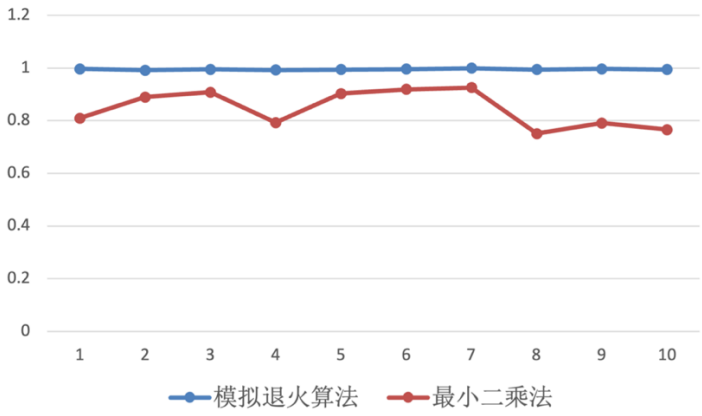


图 4 R2 决定系数对比图

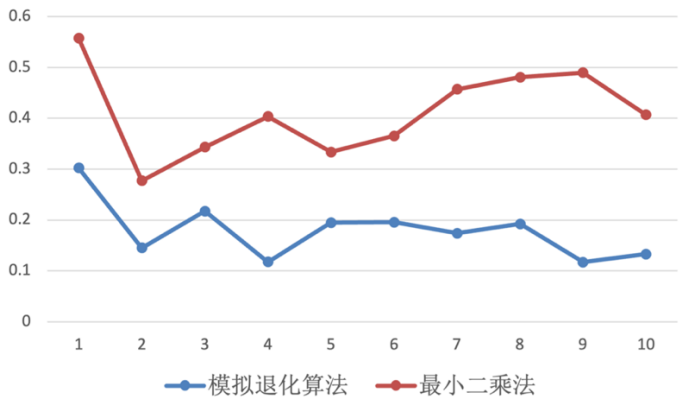


图 5 MAPE 对比图

2.5 两种算法的讨论

模拟退火算法对于参数选取基于蒙特卡洛思想，相较于拐点法对于 K 值的估计更加准确。一方面是拐点法对于数据要求较高，虽然理论上梯度最大的点应该是整个发展过程的中点，但现实场景很少有真实数据完全满足这种正态分布，而利用模拟退火得出的 K 值是从数据本身学习的，对数据分布没有要求。另一方面是最小二乘法需要先对拟合目标做对数变化，而对数的取值范始终要大于 0，因此$K-f(t)>0$，这时如果估计的 $K<max\{f(t)\}$，为保证运算正常进行就需要对 K 值作边界值处理，一般处理思路是令 K 等于$max\{f(t)\}$乘一个大于 1 的系数，这种思想不仅过于启发式，且会导致拟合的同质性高，例如很多专利都是在同一年到达成熟阶段，这很有可能是因为利用这些数据估计的 K 值均小于$max\{f(t)\}$，因此边界处理后结果相似，而模拟退火算法没有进行对数变化，因此对于 K 值的估计不需要进行边界处理，那么最终拟合结果也符合真实预期。

三、结论与未来工作

为预测在 2035 年左右可能发展成为未来产业的产业，我们收集德温特数据库中年粒度的专利数据，搭建了基于隐含狄利克雷分布和 Fisher-Pry 模型的未来产业预测模型框架。为使得模型预测准确，我们分别通过最小二乘法和模拟退火算法求解模型，并对求解结果进行对比，得出利用模拟退火算法求解未来产业预测模型效果更好的结论。值得注意的是，本文的研究是全球未来产业前沿技术预测工作的初步尝试，研究上也存在一定程度的不足之处，在模型和算法上都有待提升的空间。而随着计算机计算能力的快速提升和神经网络结构的不断发展，深度学习因其出色的数据表征能力而受到广泛关注，能否将深度神经网络耦合进未来产业预测模型框架中，这个方向研究的可能有意义的点包括，深度学习模型是否比本文所用的文本聚类模型 LDA 有更好的聚类效果， 比时序数据拟合模型 Fisher-pry 对众多不同产业的数据更有通用性等。另外的一个探索方向是使用更多种类的数据用于产业建模和预测，例如融合科技论文、科技报告、科技新闻等数据，实验更多维度的数据带来的网络效应是否可以带来模型预测能力的提高。