1. 阳光高考网站数据爬取工作
2. 学校名称和对应学校代码的爬取：

方便在代码与数据库中以int型数据表示学校。但是阳光高考网的学校代码并非是按顺序排列的，而且有空缺的地方。总之打算先按学校ID从1到2000 顺序爬取，然后找出空缺的学校，人工修补。

1. 学校各省各年各批次平均分的爬取：

用于志愿填报风险评估模块，用以估计某排名段在往年的学校平均录取情况。程序设计、与数据库连接、测试工作已经完成。

1. 模拟填报志愿风险评估模块

该模块是实现利用大数据指导志愿填报功能的核心模块，包括用户数据录入（数据库表设计）、核心代码开发。目前完成了核心代码的全部思路以及部分代码，但是由于缺少数据，还没有进行测试。

总的来说，为了评估某一考生填报志愿的风险，必须知道在用户高考排名之前的所有学生填报的志愿。这样一般来说面临着两个问题：

一是存在绝对误差，比如模拟志愿和实际志愿不一致的学生。可以考虑用一个常数来修正它。比如说总的招生计划数乘以往年平均报录比，得到模拟计划招生数。

二是存在信息残缺的问题。假设一个考生考了1000名，那么要想完全确定他能不能上南开，只需要知道

1. 前1000名学生究竟是怎么填的；（2）各学校的招生计划。

问题的关键在于前1000人里某些档位的信息我们可能没有，这时候需要系统按概率自动生成平行志愿，这就使得填报志愿能否成功有了变数。自动生成有四个来源：

（1）该排名段去年学生填报情况

（2) 该排名段今年其它考生填报情况

（3）该排名段上下相邻两段今年考生填报情况

（4）根据学校前五年线差确定的各学校风险等级

自动生成的数据是有一定的概率的，不会每次都生成同一个志愿，这就存在了变数。多次模拟这一情况，可以得到排位为m的考生的志愿成功率。

1. 实现细节

显而易见，问题的关键在于自动生成志愿部分。实现的想法分为两步，第一步是根据四个来源找出该档位合理的学校池和专业池。第二步是根据学校池得出每一种志愿组合的概率。

第一步：

1. 该排名段去年学生填报情况
2. 得出今年该排名在去年对应的分数
3. 找到分数集后，根据阳光高考网的平均分数据筛选学校

（2)  该排名段今年其它考生填报情况以及相邻两段今年考生填报情况

I. 确定该排名挡上下现有最近的两档排名

II. 得到这两档排名其它所有考生的志愿

III． 把该排名段今年其它考生的志愿添加进去

（3）根据学校前五年线差确定的各学校风险等级

I． 利用**专业推荐模块（见后）**里的函数，对学校池和专业池里的每个元素进行风险等级划分。

II． 确保生成的志愿中，高、中、低风险的学校各占一定比例。

第二步：

目前的想法还是同一风险等级无差别抓取。比如生成6个平行志愿的结果，按照常理应该是冲一冲有两个， 稳一稳有三个，保一保有一个。所以从之前已经得到的高风险池里随机选两个，中风险池里随机选三个，低风险池里随机选一个。这样，某一个学校是否出现在自动生成的志愿里的概率就可以算出来了。不过个人更倾向于对某考生前面的全部空缺档位生成很多次志愿，每一次都判断该考生是否填报成功，之后得出成功率。

1. 推荐算法的建立与检验

1.主要实现思路

为了针对考生的成绩以及排名情况做出学校推荐，首先要对报考不同学校的风险进行量化（先在只考虑能否被学校录取的基础上评估，不考虑报考专业）。对此首先建立风险量化策略。

初步评估风险按照考生排名是否超过某所学校某年招生最低排名以及超过近五年内最低排名的年数。单纯按照此种方法评估风险高低会有较大偏差。因此考虑历年录取最低排名的变化趋势，与过线年数共同作为评估风险的依据。

用*N*表示考生当年排名超过近五年内某学校最低录取分数线的年数，用表示考生排名，表示近*i*年分数线。令*r*为风险值基值，*Risk*为风险值。

当*N =* 0,1,2时，*r* = 500；

当*N =* 3,4时，*r* = 200；

当*N =* 5时，*r* = 25。

考虑到排名的数值可能非常大，因此需乘一个较小的数使之与r数量级相近。该系数又不宜过小，过小的数值起不到校正作用。经过若干次调节后选择0.03作为系数。

若近两年最低录取排名有下降趋势即:

若近两年最低录取排名波动不大（差异系数）:

若近两年最低录取排名波动较大（差异系数）:

其他情况采用五年中最高录取排名：

2.参数调整与小范围验证

为验证可用性以及调整参数，搜集了9所学校2012-2017年以前五年在河北省的最低、最高、平均录取排名，画出考生排名-风险值曲线。

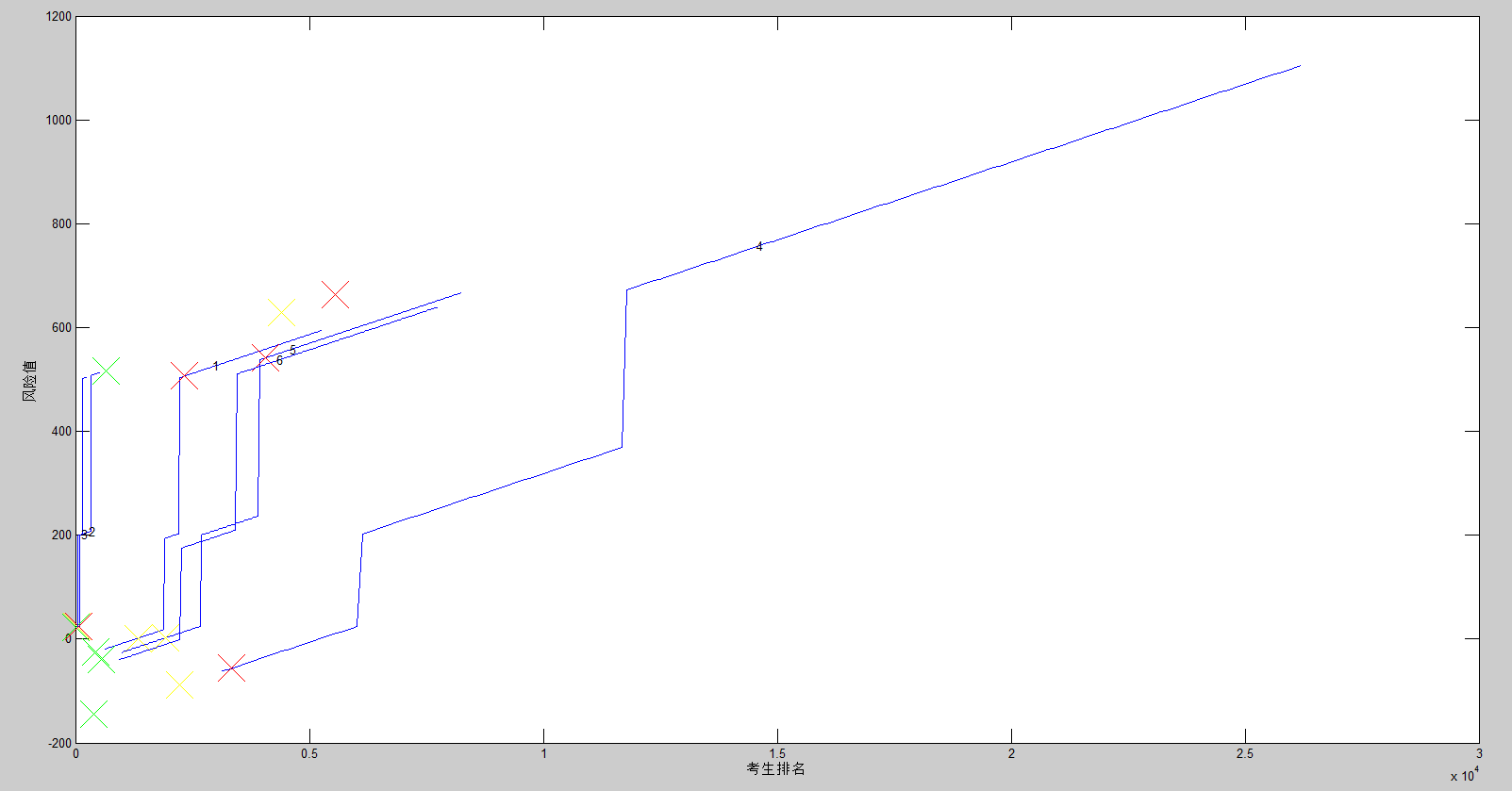


图1

图1中红点为2017年各学校最靠后录取排名，黄点为2017年各学校平均录取排名，绿点为2017年各学校最靠前录取排名。图中横坐标为考生排名，纵坐标为风险曲线，五条曲线从左到右分别为清华大学、南开大学、天津大学、哈尔滨工业大学、四川大学。

3.确定风险等级

在前面，我们得到的仅仅是风险值，这没有任何意义。我们现在要划分风险等级，让它变得有意义。选取前五年各学校最低录取排名中的最低值、最高值，并求出平均值，计算风险值，在9校风险值曲线中以横线表示出来，如图2所示。其中红色横线对应的纵坐标代表以某学校五年来最低录取分数对应排名里的最靠后排名为参数代入计算得到的风险值。黄、绿对应平均和最靠前。图中横纵坐标同上图，从左到右的九个学校分别是清华大学、复旦大学、南开大学、天津大学、哈尔滨工业大学、四川农业大学、石河子大学、广东海洋大学。最后得到的结果表明，尽管学校差异巨大，但是这些横线明显分成了三族，分别集中在风险值等于500、200、22.7的位置。（我认为，即便取更多的数据，最后依然会集中于这个位置附近）

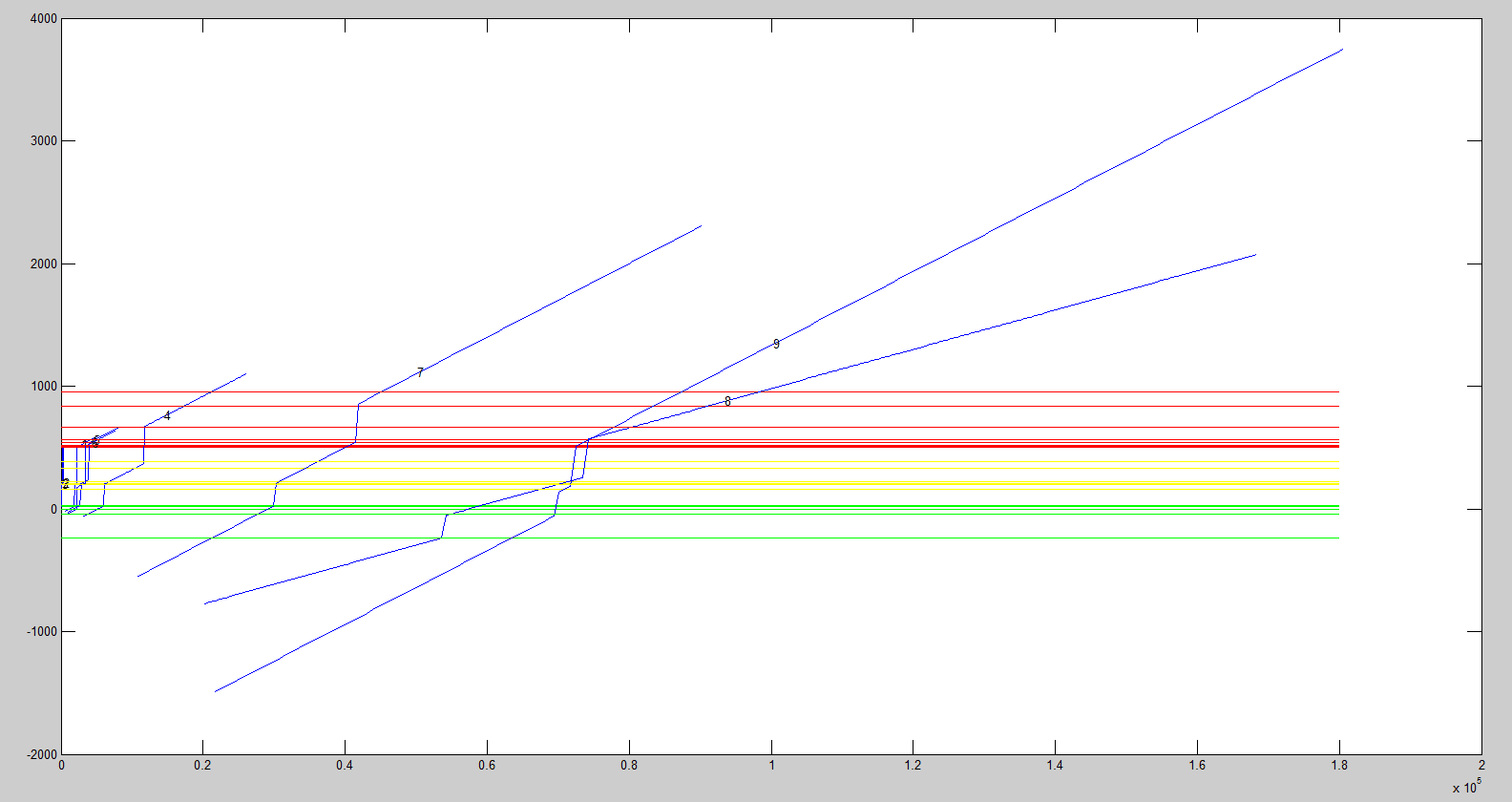


图2

以此为依据，取三个集中值中间的两个位置为风险等级分界线，即，风险值小于111.35的为低风险，风险值在111.35到350之间的是中等风险，风险值高于350的排名为高风险。

4.对风险等级的检验和推荐列表的生成

我们找了2017年的5所学校录取的最高分、最低分和平均分，点的数字1-5分别是南开大学、清华大学、四川大学、天津大学和哈工大。作为对预测结果的验证。图上的点表示的是以17年的各学校录取排名里的最靠后的（红色）、最靠前的（绿色）和平均（黄色）对应的风险值。结果表明，17年的录取最低分基本上在高风险和中等风险的交界处，平均分和最高分则几乎在低风险和中等风险区域。

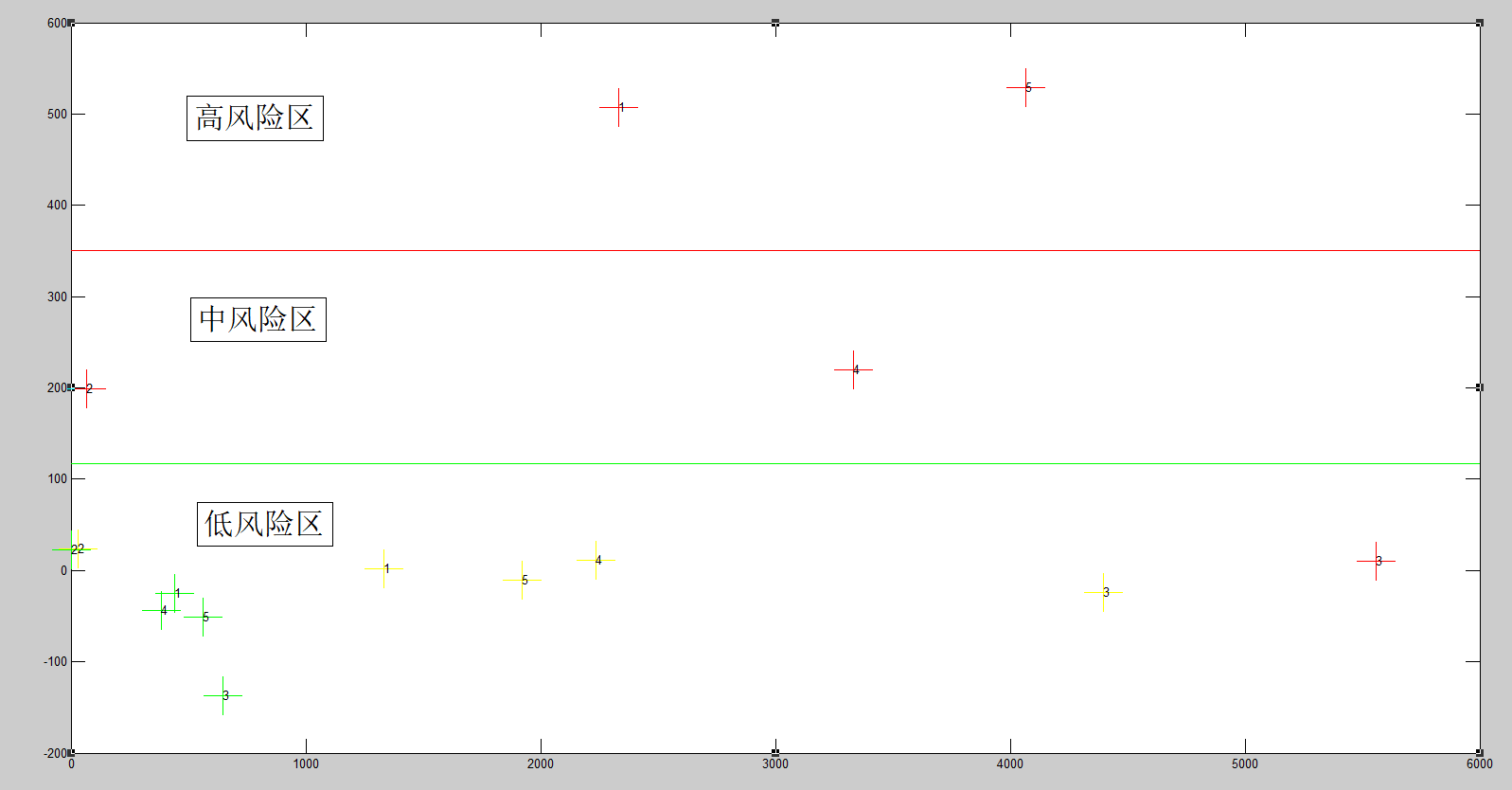


图3

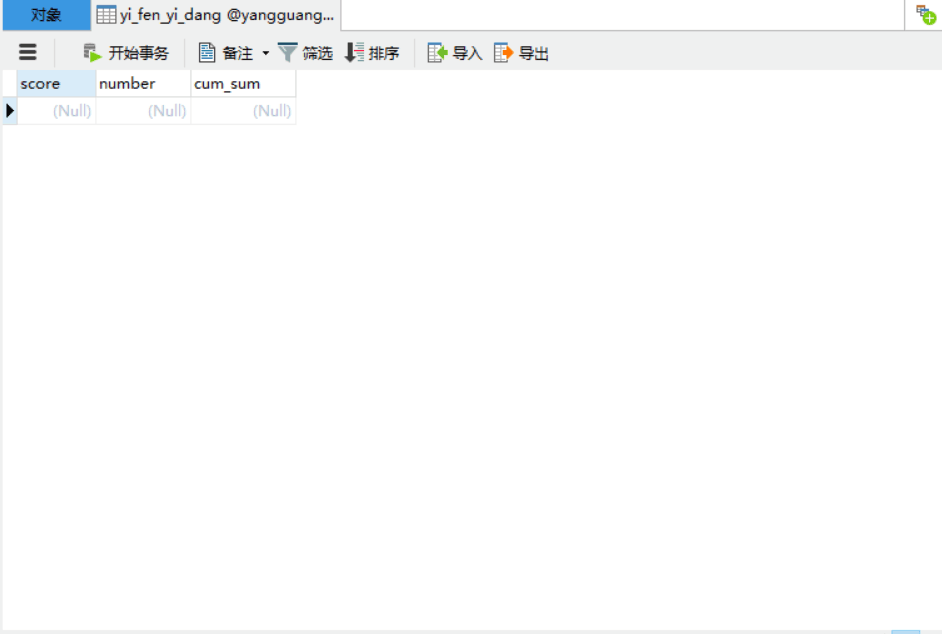
由图可见，除四川大学外结果都比较好，原因为此五年四川大学的最低录取排名相当不稳定，从6000名到17000名都有，同时最高录取排名也存在某些年份非常高的情况。如此看来，作为初步生成推荐列表的依据，此种风险分法较为可靠。

生成推荐列表时，综合学科评估结果，从低风险中选择5所学校，高中风险共选择5所学校，按照学校排名以及风险高低列出初步推荐列表。

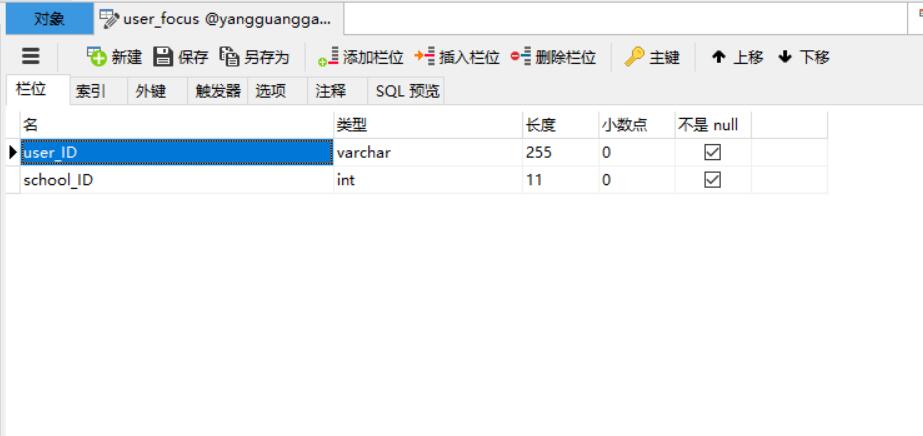
四、数据表结构的初步建立



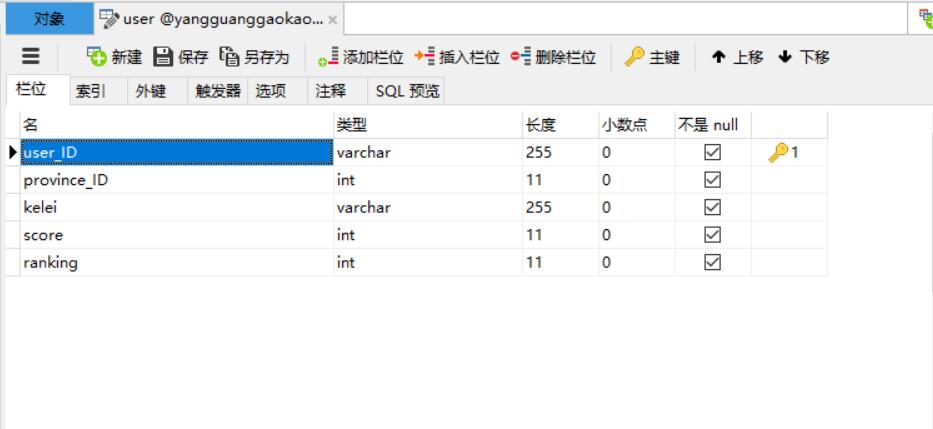
初步建立以上七个表，因为学校招生信息结构较为复杂，因此分年份存储。



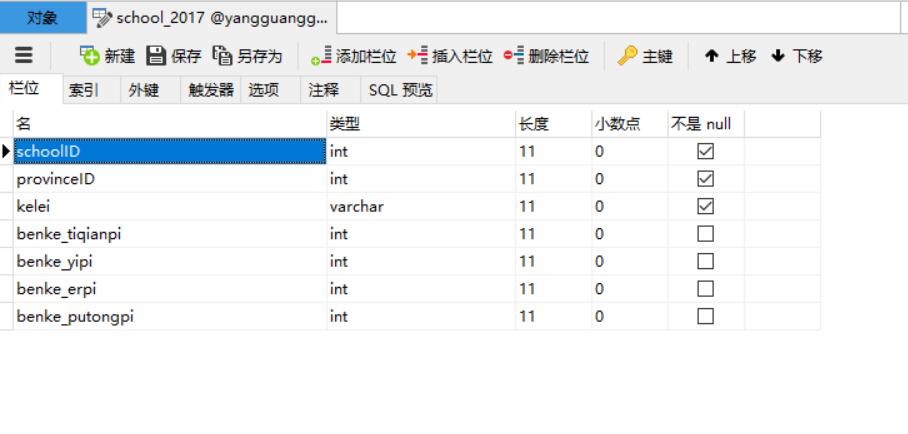
一分一档表



关注关系表



用户信息表



学校信息表（以2017年为例）

数据量较大，结构也较复杂，今后可能还需要随算法继续设计进一步调整。