**8个基础的特征向量：**

**1.性别 2. 年龄 3.招聘渠道 4.文凭 5.在职年龄 6. 婚姻状况 7. 健康状况 8. 职称**

**使用的算法：RBF核的SVM**

**加强部分：将测试数据分为多组，构建多个模型，在新数据进来的时候，用多个模型预测结果。产生的结果再进行投票，产生最终的结果。**

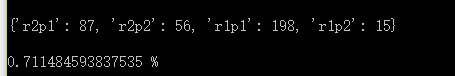
**实验验证——提高胜率**

**一．对模型的提升**

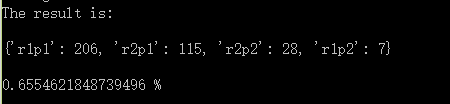
1. ***测试模型的数据数量与准确率关系***

**A之前：1000个数据单独拆开成4组，每组250个, 8个基础特征向量**

8+随机分布，默认参数：

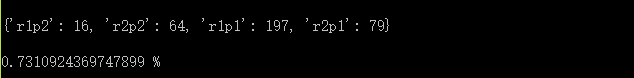


8+随机分布，调整参数：

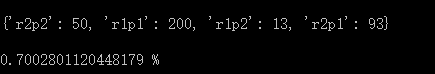


**B新:四组，第一组1-500, 第二组151-650, 第三组301-800, 第四451-950**

8+随机分布，默认参数：



8+随机分布，调整参数：



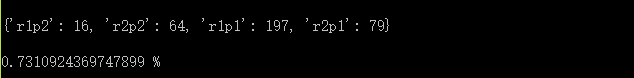
**结论：在此种模型下，提高测试模型的数量在其余条件相同的情况下对结果由显著的提升**

**原因：产生数学模型的数据更多，因此在测试新数据的时候能够更加准确的进行分类**

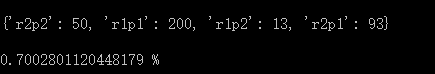
***2.提高参与投票的模型的组数***

**A之前：1000个数据单独拆开成4组，每组500个, 8个基础特征向量。第一组1-500, 第二组151-650, 第三组301-800, 第四451-950**

8+随机分布，默认参数：



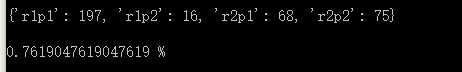
8+随机分布，调整参数：



**B新:11组，第一组1-500, 第二组51-550,101-600**

**缺点：由于需要对11组数据进行统计然后投票，程序运行速度明显降低**

8+随机分布，默认参数：



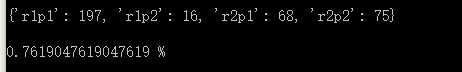
**结论：当提高模型数量时，由于采用了bagging的思想，有投票机制，预测的准确率得到了一个显著的提升**

**原因：模型较多，且比较全，利用了几乎所有的数据，在投票的时候，更容易获得较为理想的结果**

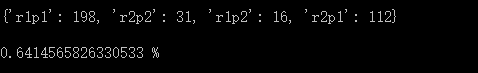
***3.数据进行归一化后影响:***

**A之前：**

8+随机分布，默认参数：



**B新的测试胜率：**



**结论：当对所有数据都做一个scale扫描之后，模型的正确率有着较大的下降。原因：从归一化的数据进行分析，由于归一化后，比较重要的项例如在职时常被归一化到了(-1,0)之间（之前有相对较大的数值差距）；而其他不那么重要的数据却归一化的区间是(-1,1)之间（之前数值差距都只有1），区间间距变化，获得权重变强。导致了正确率的下降**

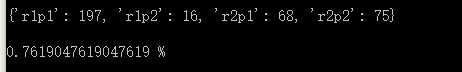
**二．对特征向量提升（在没归一化情况下）**

**在之前11组做投票的基础上，改变特征向量，验证最后胜率**

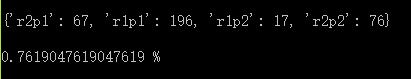
**1.性别 2. 年龄 3.招聘渠道 4.文凭 5.在职年龄 6. 婚姻状况 7. 健康状况 8. 职称**

1. ***减少性别特征向量***

**A之前的胜率：**



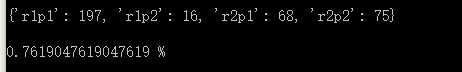
**B新的测试胜率：**



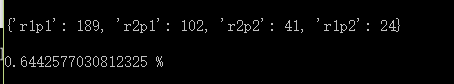
**结论：从结果进行分析，混淆矩阵只发生了一个变化，多预测准了一个会留下来的，少预测了一个会走的。最后结果并没有发生变化，性别并不是影响离职的重要因素**

1. ***减少年龄特征向量***

**A之前的胜率：**



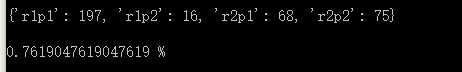
**B新的测试胜率：**



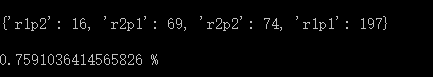
**结论：年龄对是否离职有着相对比较大的影响**

1. ***减少招聘渠道***

**A之前的胜率：**



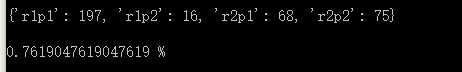
**B新的测试胜率：**



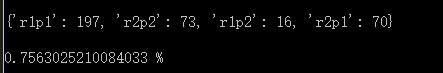
**结论：招聘渠道对是否离职有着较小的影响**

1. ***减少文凭特征向量***

**A之前的胜率：**

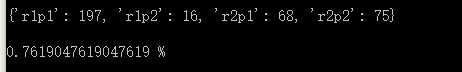


**B新的测试胜率：**

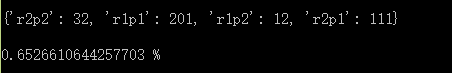


***5.减少在职年龄特征向量***

**A之前的胜率：**

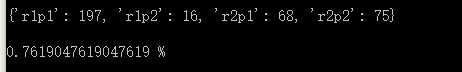


**B新的测试胜率：**

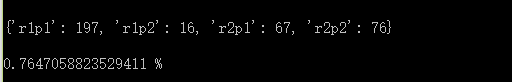


***6.减少在婚姻状况***

**A之前的胜率：**



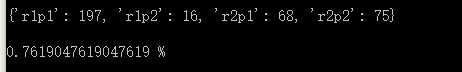
**B新的测试胜率：**



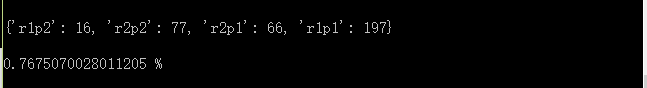
**结论：第一次出现提升，说明婚姻状况在此判断中起到了反效果**

***7.减少在健康状况***

**A之前的胜率：**

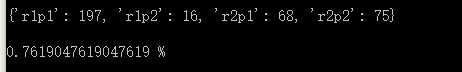


**B新的测试胜率：**

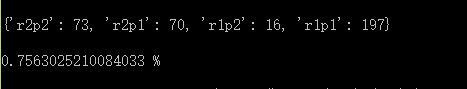


***8.减少在职称***

**A之前的胜率：**

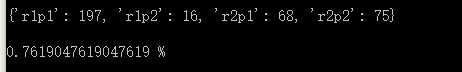


**B新的测试胜率：**

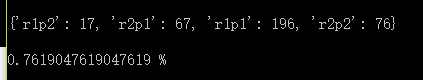


***8.减少婚姻状况+健康状况***

**A之前的胜率：**



**B新的测试胜率：**



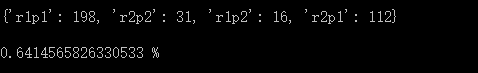
**三．对特征向量提升（在数据归一化情况下）**

**在之前11组做投票的基础上，改变特征向量，验证最后胜率**

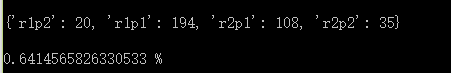
**1.性别 2. 年龄 3.招聘渠道 4.文凭 5.在职年龄 6. 婚姻状况 7. 健康状况 8. 职称**

***1.减少性别特征向量***

**A之前的胜率：**

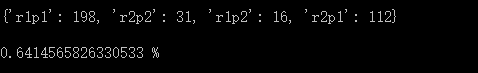


**B新的测试胜率：**

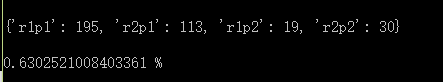


***2.减少年龄特征向量***

**A之前的胜率：**

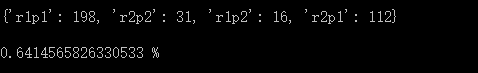


**B新的测试胜率：**

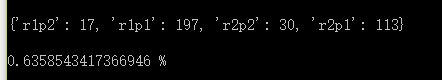


***3.减少招聘渠道***

**A之前的胜率：**

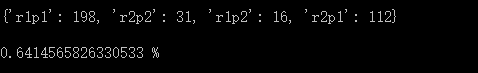


**B新的测试胜率：**

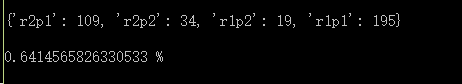


***4.减少文凭特征向量***

**A之前的胜率：**

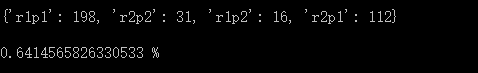


**B新的测试胜率：**

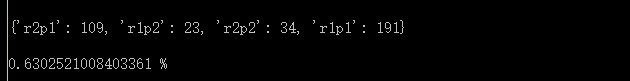


***5.减少在职年龄特征向量***

**A之前的胜率：**



**B新的测试胜率：**



**结论：在对数据进行归一化后，前五个特征向量都对结果影响不大，结果很奇怪**

**四．特征向量提升**

**之前对不同的类别进行了人为干涉的数学转化，当出现一个新类别的时候，在原有的类别数字上加1，例如：招聘类型中有客户推荐、网站招聘、应届招聘等，在转化过程中客户推荐---->1、网站招聘---->2、应届招聘----->3依次类推。由于每个数值之间只差1，导致的结果是间隔过小，可能对最后分类结果形成影响。所以决定做一个探究实验。**