数据挖掘与知识发现

期末综合实践

姓　名：李品鑫

学　号：20191060239

学　院：信息学院

专　业：智能科学与技术

年　级：2019 级

2022 年 6 月 1 日

对员工离职数据集进行处理并利用贝叶斯分类器建立的企业员工流失预警模型

【实验介绍】:

本次实验研究了一个企业的员工离职数据集，通过建立一个贝叶斯分类器完成员工流失预警模型，经过在数据集上划分的数个不同测试集测试表明，该分类器的准确率稳定在 74%~76%之间，足以证明这个分类器可以对员工离职作出准确有效的预警。

【实验背景】:

企业培养人才需要大量的成本，为了防止人才流失，对员工流失因素的分析尤为重要。员工流失分析是评估公司员工流动率的过程，目的是找到影响员工流失的主要因素，预测未来的员工离职状况，以及时地做出改变与调整，对人才的流失进行预防。

【实验任务】

基于提供的员工离职数据集，数据主要包括影响员工离职的各种因素（如员工满意度、绩效考核、参与项目数、平均每月工作时长、工作年限、是否发生过工作差错、5 年内是否升职、部门、薪资）以及员工是否已经离职的对应记录，基于本学期所学知识构建员工流失预警模型，并为企业避免员工流失提供建议。

【实验过程】:

1. 数据分析

该数据共有 14999 条员工记录，每条记录中包含着 11 个特征。刨去分类中无用的编号与代表离职状况的特征，我们可以得出以下 9 个有用的特征来构成特征向量，它们的名称与所代表的特征分别为：

属性	satisfaction_level	time_spend_company	average_montly_hours	last_evaluation
类别	连续，数值	连续，数值	连续，数值	连续，数值

属性	work_accident	number_project	promotion_last_5years	sector	salary
类别	离散，逻辑	连续，数值	离散，逻辑	离散，字符	离散，字符

可见，该数据集的特征维度高，数据类型复杂，并且含有缺失值。故在进行分类预测之前，我们需要先对数据集进行预处理，包括空值处理、连续数据离散化等。

2. 数据预处理

2.1. 缺失值处理

在 Excel 表格中对缺失值 ”NULL” 进行查找计数，统计发现在 14999 条样本数据中仅有 11 条样本包含缺失值，所占比例极小，故直接删除含缺失值的样本，即可，缺失值处理后的数据集包含 14988 条数据。

2.2. 离散化处理

数据的离散化是将连续型数据划分为数个离散性数据的过程。通过对数据的观察，决定对 satisfaction _level、average_montly_hours、last_evaluation 三个特征进行离散化处理。至于同为连续型数据的 time_spend_company、number_project，因其为整数型，且数值范围小，故在分类时可以直接将其看作是离散数据。

离散化处理前，我们要先观察数据的概率分布特征。

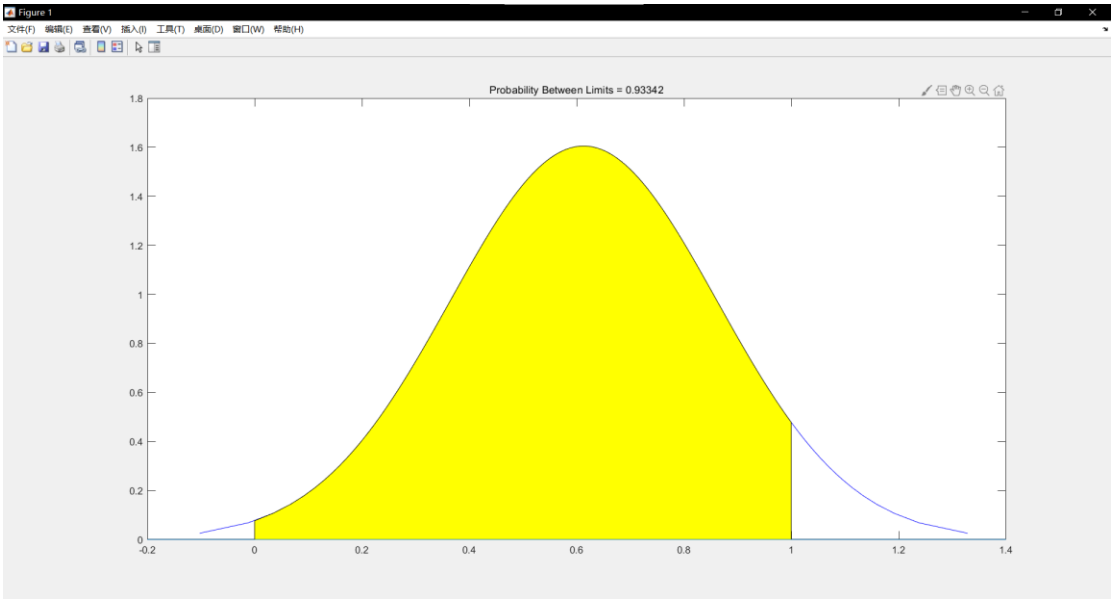

图 2.2.1- satisfaction _level 的分布曲线

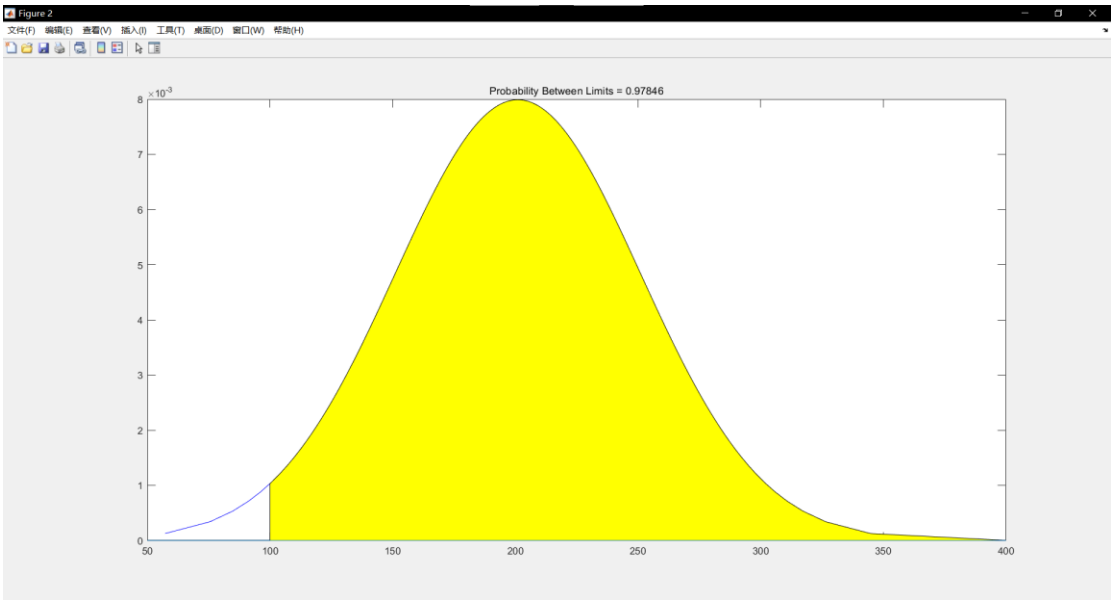

图 2.2.2- average_montly_hours 的分布曲线

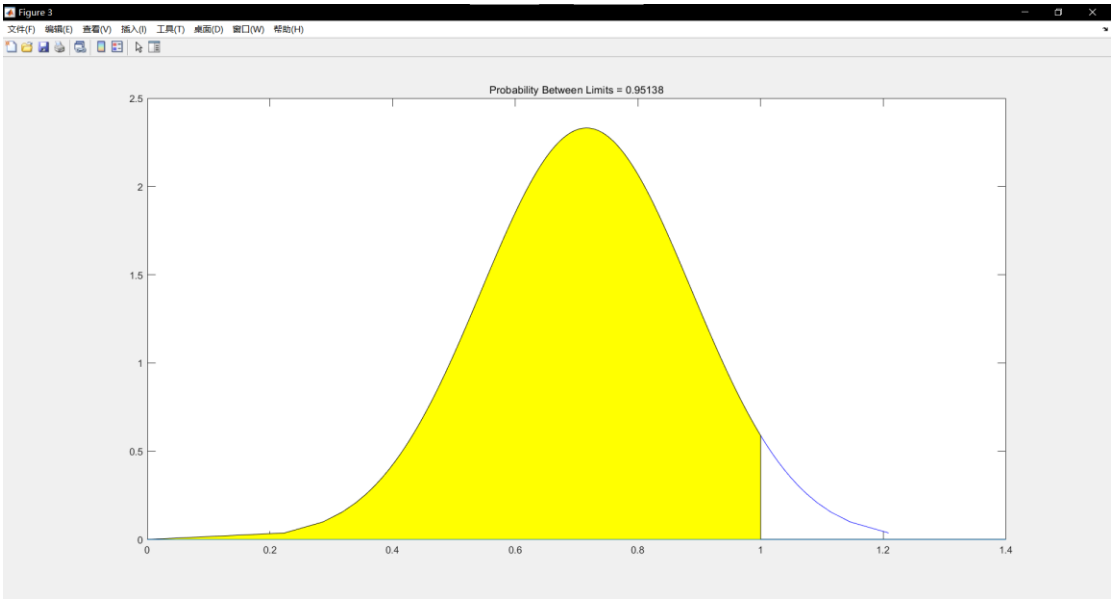

图 2.2.3- last_evaluation 的分布曲线

可以看出，以上三个特征都遵从一定的分布模型，为保持其分布特征不变，故采用等宽法划分为离散化区间。

3. 构造预警模型以及测试

3.1. 采用的算法介绍

本分类其采用了朴素贝叶斯分类算法，朴素贝叶斯算法采用了贝叶斯公式：

$$P(y_i|x_1 \dots x_n) = \frac{P(y_i)P(x_1 \dots x_n|y)}{P(x_1 \dots x_n)}$$

其中 $y = \{y_1 \dots y_m\}$ 为分类的类别， $x = \{x_1 \dots x_n\}$ 为一个待分类项，x_i为特征向量的分量。因$P(x_1 \cdots x_n)$在计算中可被视为常数，故而可将公式简化为：

$$P(y_i|x_1 \dots x_n) = P(y_i)P(x_1 \dots x_n|y)$$

利用该公式依次计算待判别样本属于全部分类类别的概率值，得到其中最大的值

$$Max\big(P(y_i|x_1 \dots x_n)\big), \qquad \mathrm{i} = \{1 \dots \mathrm{m}\}$$

在朴素贝叶斯算法中，最大概率值所对应的类别即为分类结果。

在本分类器中，数据被随机分为互斥的训练集与测试集，它们分别占有总数据集 60%与 40%的比重。对该数据集进行多次的随机划分与分类实验，避免小概率意外的发生。

3.2. 分析与评估

使用该分类器针对员工离职数据集进行分类，进行 10 次重复实验，其正确率如下表：

次数	1	2	3	4	5	6	7	8	9	10
总个数	5995									
正确个数	4503	4582	4437	4461	4505	4526	4582	4530	4591	4511
正确率	75.11%	76.43%	74.01%	74.41%	75.15%	75.50%	76.43%	75.56%	76.58%	75.25%

数次运行都得出了一个比较稳定的结果，足以证明这个分类器可以对员工离职作出准确有效的预警。

4. 员工流失因素分析与建议

4.1. 流失因素分析

为评价特征向量的各分量对员工离职的影响，我们可以以后验概率作为评价指标，即“在离职的员工中有$P(x_i|y=1)$ 的员工拥有特征x_i”，将后验概率大于一定阈值的特征分量视为对员工离职有着有较大影响的因素。设定最大阈值

$max\_con = 0.3$，进行查找，查找结果如下表：

特征分量	satisfaction_level	time_spend_company	average_montly_hours	last_evaluation
值	1, 2, 3（对应离散化之前的 0~0.6 区间）	3,4,5	2, 5（对应离散化之前的 100~150, 250 以上间）	3, 5

特征分量	work_accident	number_project	promotion_last_5years	sector	salary
值	0	0	0	sales	low, medium

由上表可总结出以下几点：

（1）员工满意度到 0.6 以下的员工更倾向于离职

（2）工作年限为 3~5 年的员工更倾向于离职

（3）平均月工作时间在 100~150 与 250 小时以上的员工更倾向于离职

（4）绩效考核低于 0.6 与高于 0.8 的员工更倾向于离职

（5）大部分离职的员工并未发生工作事故

（6）大部分离职的员工未参与项目

（7）最近 5 年未升职的员工更倾向于离职

（8）大部分离职的员工为销售职位

（9）工资为低或者中的员工更倾向于离职

因此，为了避免人才流失，公司需注意：

（1）提高员工满意度

（2） 为平均工作时长低的员工分配更多任务，同时降低工作时长高的员工的工作负担。

（3）提供更多的升职渠道与机会

（4）提高工资

5. 实验小结

本次实验采用了朴素贝叶斯分类器，仅以计算得出的最大权值作为评价指标。且因维度较大，运行速度较慢。经过反思后我找出了以下几个改进点

（1）可进行数据降维

在机器学习和统计学领域，降维是指在某些限定条件下，降低随机变量个数，得到一组“不相关”主变量的过程。在该数据集中先应用降维处理后再进行贝叶斯分类可以加快运行速度。

（2）未引入风险权重

不同员工对公司的重要度是不同的,（例如一个采购岗位离职后重新招聘替补人员的难度不大，但是一个技术人员或者是高级管理人员离职后想要再找到这样的人才就难了）。因此，根据不同员工的重要性主观地去设置一个“权重”来预估是否会离职以及是否要对其进行干预。

（3）无法分析相关性

贝叶斯分类器无法找出数据中的相关性,（例如大部分离职员工都是低薪水，于是贝叶斯分类器认为中等薪水对离职带来的影响十分小。然而，公司中占大多数的低级别员工更多满足于中等薪水于是选择离职，但是占少数却十分重要的高级员工对中等级别薪水觉得不满，于是选择离开）。

（4）当不同类别的数量差距悬殊时，贝叶斯分类器将会失去效果。

本实验采用的分类器起初是按照$P(y_i|x_1 \dots x_n) = P(y_i)P(x_1 \dots x_n|y)$ 的公式进行计算的，然而在计算过程中发现，由于$P(y_i)$的差距巨大$P(left{:}\,0) = 76.23\%$ ， 而 $P(left{:}\,1) = 23.77\%$ ， 过大的差距严重影响了$Max\big(P(y_i|x_1 \dots x_n)\big),\ \mathrm{i} = \{1 \dots \mathrm{m}\}$的计算，导致最初的分类正确率仅有

20%~30%。因此采用了消去$P(y_i)$后的分类公式$P(y_i|x_1 \dots x_n) = P(x_1 \dots x_n|y)$，实践证明该公式正确率达到了 75%，是有效的。