上海大学　2019 — 2020 学年　冬季学期

课程名称：机器学习与力学

课 程 号：01826249

授课教师：胡国辉

论文名称：快艇水动力参数朴素贝叶斯概率计算

学　　号：17121632

姓　　名：翟晗锋

专　　业：理论与应用力学

成　　绩：

日　　期：2020/3/10

评　　语：

快艇水动力参数朴素贝叶斯概率计算

翟晗锋

（上海大学力学与工程科学学院，上海 200444）

摘 要

快艇在当今社会已经非常常见。快艇在航行中会有会有多种复杂的参数指标。在实验室中相关的研究分析并记录下不同属性的数值可以帮助我们了解快艇动力参数如何控制快艇运动。由代尔夫特理工大学海事和运输技术系船舶流体力学实验室提供的游艇水动力数据集中数据进行朴素贝叶斯算法计算，我们可以利用概率计算结果判定在特定设置条件下快艇的残余阻力大小。因为数据为实验室定量过程测算，顾其结果为连续的。为方便进行朴素贝叶斯运算，我采用间断值分类的方法将数据组分为 3 级进行概率计算。根据相关结果，我们还可以分析何种条件为快艇航行的最佳条件，即阻力最小的情况。利用这些分析计算结果，可帮助控制快艇运行条件，同时更好求解流体动力学相关问题。

关键词： 流体力学；水动力参数；朴素贝叶斯；机器学习

1 引言

快艇是当今人们休闲娱乐生活中必不可缺的一部分。随着生活质量不断提高，越来越多的人开始享受快艇娱乐带来的快乐体验。快艇最吸引人的地方莫过于其高速航行速度和强劲的动力，这就需要其良好的流体动力学参数。对于一名力学学生来说，研究快艇的动力学参数对于我们在流体力学方面的学习会有很大的帮助和提升。同时，结合我们在课堂所学的朴素贝叶斯分类方法，对快艇给定参数下进行残余阻力的概率计算，我们可以预测出快艇在给定条件下更倾向于出现多少的残余阻力，即出现怎样的运动。这样就可以帮助我们判断快艇高速航行所需要控制的条件。结合流体动力学方面的数据与机器学习中朴素贝叶斯分类方法，我们可以将一些传统难以求解的问题简单化、准确化。

2 背景概括

2.1 快艇

快艇（或游艇）一类用于娱乐，巡航或竞赛的帆或动力船。许多赛艇都是经过特殊设计的船只，其起居空间经设计达到最少，重量更轻。同时，还有一些游艇，即是当今社会人们熟知的娱乐游艇，是为牟利而经营的。

游艇长度通常在 7 米到数十米的范围内。小于 12 米的动力艇，可以用来过夜居住。超过 24 m 的游艇可被归类为“大型”游艇，适用更高的建造标准；“商业”游艇可携带不超过 12 名乘客；“私人”游艇仅是为了船东和客人的娱乐，或由标志，国家下其注册。

图1 由D Ramey Logan拍摄的被称为“炼金术士”80英尺游艇

Fig.1. 80 foot motor yacht Alchemist photo D Ramey Logan.

2.2 流体动力学

流体动力学是在流体力学中分析流体动力系统作用分析的研究。其中，在我们分析这样问题时，需要分析运动流体属性主要为流体速度、压力、密度、温度等。

流体动力学应用广泛，例如预测天气，计算航天器如飞机受力&力矩，输油管线中石油的流率等方面上，甚至血管中血液流动问题。甚者可运用在交通工程方面，这时交通运输中可以应用连续流体假设来分析。还包括了解星云在星际空间和造型裂变武器引爆等。

流体动力学提供了一个系统的结构，它是这些实践学科的基础，它包含从流量测量得出的用于解决实际问题的经验和半经验定律。解决流体动力学问题的方法通常涉及计算流体的各种特性，例如流速，压力，密度和温度，它们是空间和时间的函数。

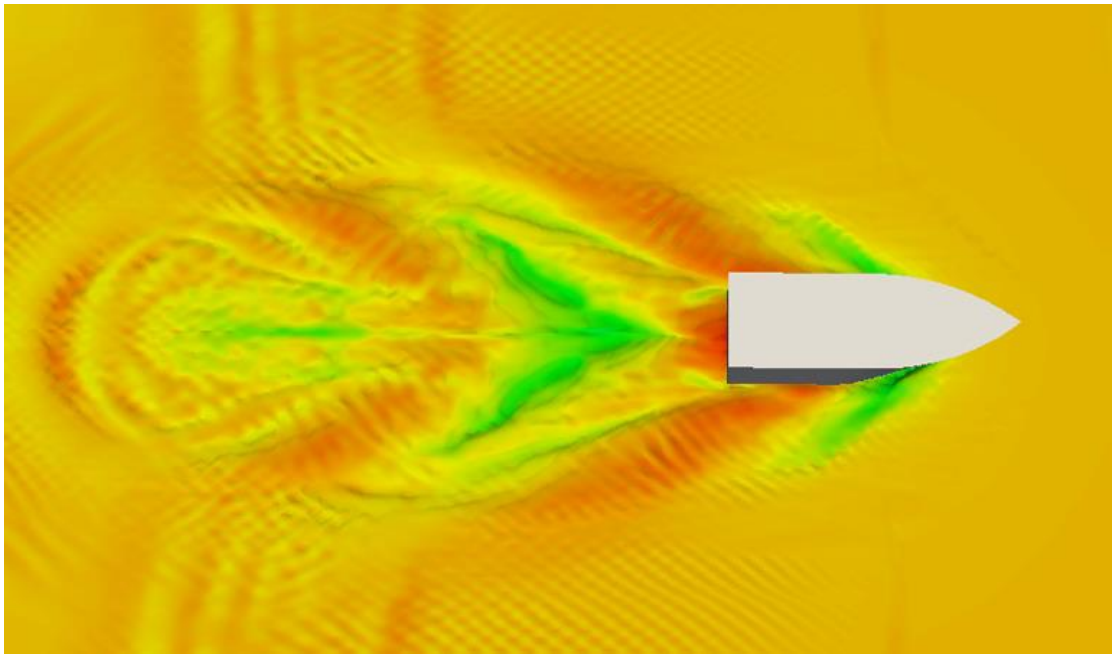

图 2 快艇在水中航行有限元模拟结果

Fig.2. Finite element analysis of the ship sailing in water.

3 方法

3.1 概念及定义

假设存在 N 种可能的类别标记，即 $\Upsilon = \{c_1, c_2, \ldots, c_N\}$, λ_{ij} 是将数据库中属性 c_j 样本错误分类为 c_i 之损失。而我们引入后验概率 $P(c_i|\mathbf{x})$可得使样本 $\mathbf{x}$ 分类为 c_i 产生期望损失，定义为样本 x 上的“条件风险”

$$R(c_i|x) = \sum_{j=1}^{N} \lambda_{ij} P(c_j|\boldsymbol{x})$$

目的为寻找一个判定准则 h：$X \rightarrow \Upsilon$ 以最小化总体风险

$$R(h) = E_x[R(h(\boldsymbol{x})|\boldsymbol{x})]$$

分析得，对逐样本 $\mathbf{x}$，若 h 让条件风险 $R(h(\mathbf{x})|\mathbf{x})$取得最小值，那么总体风险 $R(h)$也取得最小值。此即贝叶斯判定准则：让总体风险取最小，即分别在逐样本选择让条件风险 $R(c|\mathbf{x})$取得最小值的类别标记，即

$$h^*(\boldsymbol{x}) = \underset{c \in \Upsilon}{\operatorname{argmin}} R(c|\boldsymbol{x})$$

由上式，h^*就被定义是贝叶斯最优分类器，相应的总体风险 $R(h^*)$我们定义为贝叶斯风险。$1-R(h^*)$对应该算法最高的计算精度，即通过算法产生模型所能达到的精度理论值最高。

具体来说，若目标是最小化分类错误率，则误判损失 λ_{ij} 可写为

$$\lambda_{ij} = \begin{cases} 0, & \text{if } i = j; \\ 1, & \text{otherwise.} \end{cases}$$

此时条件风险

$$R(c|x) = 1 - P(c|\boldsymbol{x})$$

于是，最小化分类错误率的贝叶斯最优分类器为

$$h^*(x) = \underset{c \in Y}{\operatorname{argmax}} P(c|\boldsymbol{x})$$

即对每个样本 $\mathbf{x}$，选择能使后验概率 $P(c|\mathbf{x})$最大的类别标记。

同时，要让贝叶斯判定准则来让决策风险最小化或预测结果，首先要获得后验概率 $P(c|\mathbf{x})$.然而，在现实任务中该量难以直接获得。该算法的重要任务就是估计出后验概率 $P(c|\mathbf{x})$. 总的，主要可实施 2 种方法：给定 $\mathbf{x}$，再使直接建模 $P(c|\mathbf{x})$以此预测 c，定义为“判别式模型”；也可以先对于联合概率分布 $P(\mathbf{x}, c)$建模，再次获得 $P(c|\mathbf{x})$;

$$P(c|\boldsymbol{x}) = \frac{P(\boldsymbol{x}, c)}{P(\boldsymbol{x})}$$

由贝叶斯定理，$P(c|\mathbf{x})$可写为

$$P(c|\boldsymbol{x}) = \frac{P(c)P(\boldsymbol{x}|c)}{P(\boldsymbol{x})}$$

上式中，$P(c)$被定义为类“先验”概率；$P(\mathbf{x}|c)$即样本 $\mathbf{x}$ 对应类标记 c 的形似条件概率；$P(\mathbf{x})$为整合后的取值。对给定样本 $\mathbf{x}$，定义整合的因子取值 $P(\mathbf{x})$和类标记没有关联，所以概率 $P(c|\mathbf{x})$的计算过程就可以看作如何根据给定训练数据 D 来估计先验概率 $P(c)$和最终似然概率 $P(\mathbf{x}|c)$.

3.2 问题及数据分析

在该问题中，我们先需对数据集合变量进行分析。在数据集中，总共有 6 组自变量，而我们需要判断的量为第 7 组量。在该问题中需要分析判断的的量在原数据组英文名称及其相关中文翻译如下表中所示。

其中，从物理层面分析，快艇航行快慢的分析量可以通过第 7 个参量[残余阻力]来判断。而我们在本次分析中的任务即是分析在给定条件下(确定的[1]~[6]的属性值的取值)条件下第 7 个参量的不同的取值概率。

[1] Longitudinal position of the center of buoyancy, adimensional.		
[2] Prismatic coefficient, adimensional.		
[3] Length-displacement ratio, adimensional.		
[4] Beam-draught ratio, adimensional.		
[5] Length-beam ratio, adimensional.		
[6] Froude number, adimensional.		
[7] Residuary resistance per unit weight of displacement, adimensional.		
[1] 无维数浮力中心纵向位置	**[2]** 无量纲棱镜系数	**[3]** 无尺寸长度-位移比
[4] 无维数束流比	**[5]** 无尺寸长光束比	**[6]** 无量纲弗洛伊德数
[7] 三维情况下每单位位移质量的残余阻力		

表 1 数据集中的中英文对照不同属性名称及序号定义

Table 1. Different attributes and definition on the attributes' numbers in both English and Chinese version in database.

同时，在这 7 组变量中，下载所得数据中这 7 个参量的数据值是连续的。为方便分析问题，我们需要将连续的数据化为不同的间断值。其中，将连续数据分类的判定方法与标准如下表所示。

分类值 / 属性	**1**	**2**	**3**
[1]	-5	[-2.4，2.2]	0
[2]	(-∞，0.55]	(0.55, 0.568)	[0.568, +∞)
[3]	[4.34, 4.5)	(4.5, 5)	(5, +∞)
[4]	(-∞，4]	(4, 5)	[5, +∞)
[5]	(-∞，3)	[3, 3.5)	[3.5, +∞)
[6]	(-∞，0.2)	[0.2, 0.3)	[0.3, +∞)
[7]	(-∞，1]	(1, 10]	(10, +∞)

表 2 对于数据集中连续数据进行分类

Table 2. The classification on continuum data in database.

分类后，为了判定分类的结果科学性，需要对分类前后的数据图表分析。

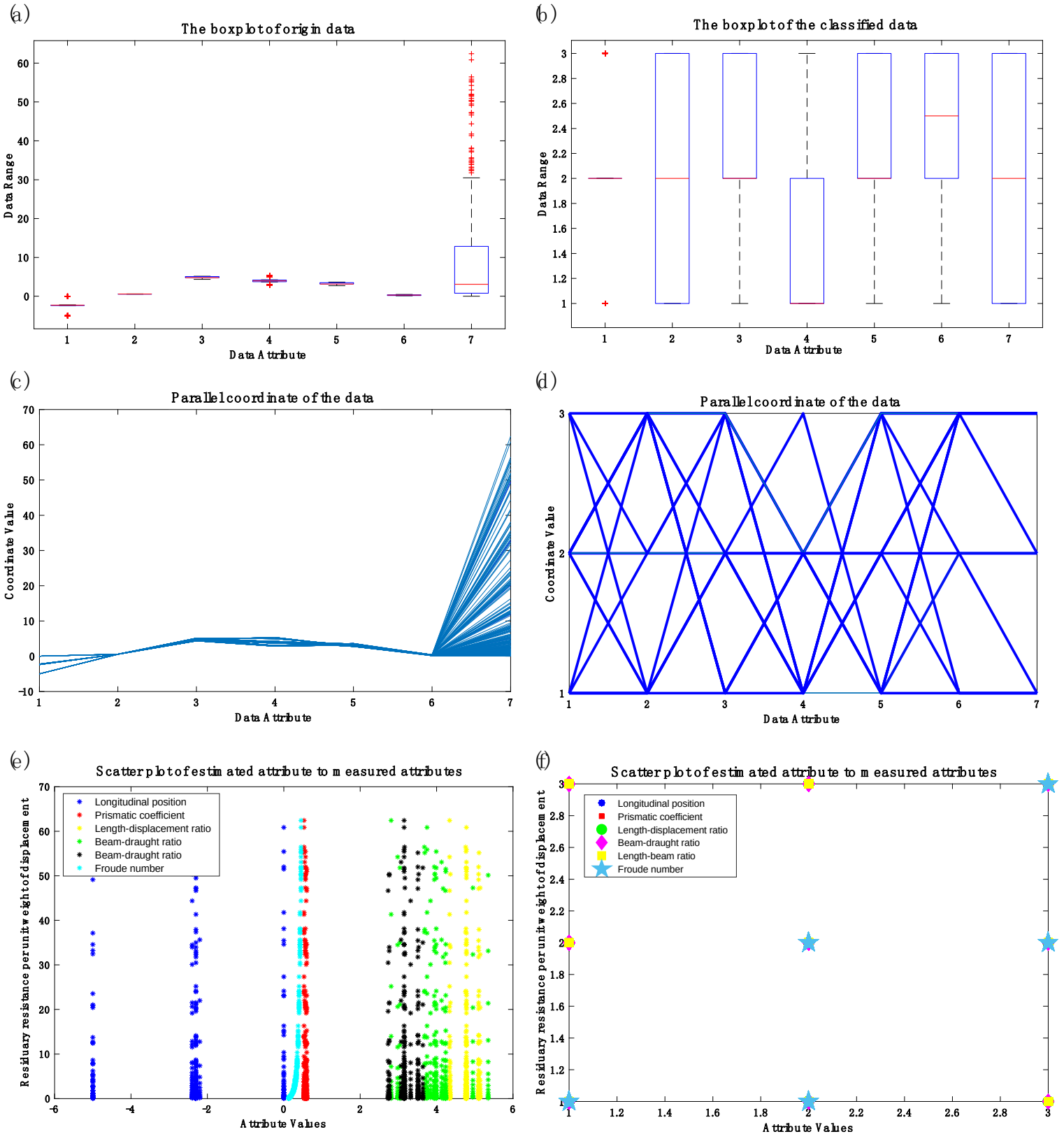

图 3 针对属性数据分类前后数据进行作图。(a)分类前数据四分位数图。(b)分类后数据四分位数图。(c)分类前数据可视化。(d)分类后数据可视化。(e)分类前数据散点分布图。(f)分类后数据散点分布图。

Fig.3. Plotted figure of data before and after classifying the attribute data. (a) A boxplot of the data before classification. (b) A boxplot of the data after classification.

(c) Data visualization before classification. (d) Data visualization after classification. (e) Scatter plot of data before classification. (f) Scatter plot of data after classification.

根据图中(a)所示，我们可以观察，分类前，这 7 组属性数据分布较为分散。同时，对每一组数据，数据的散点分布也不同。在分类后(b)，数据的中位数、平均值、被重新进行分布。(c)数据可视化可以观察 7 个快艇动力学属性数据值的关联性和分布的关系。分类后(d)在三个取值下[1, 2, 3]的关联性很清晰地被展示出来。(e)快艇流体动力学数据散点分布，7 个颜色分别对应 7 个属性；(f)对连续数据间断分类后，属性值对应的颜色数据的散点分布。

对间断分类后的数据，我们分别计算 6 个动力学属性([1], [2],…[6])中的取值为[1, 2, 3]所对应的每个属性中的概率. 其中每个取值的概率如下表所示。

检测属性数值	1	2	3
[1]	0.181818	0.636364	0.181818
[2]	0.272727	0.318182	0.409091
[3]	0.181818	0.545455	0.272727
[4]	0.681818	0.272727	0.045455
[5]	0.136364	0.590909	0.272727
[6]	0.214286	0.285714	0.5

表 3 对于分类后的数据进行概率计算。

Table 3. Probability calculation for classified data.

根据上表中计算结果所示，我们可以分别判断出对于每个单独的属性，那一个属性的概率值最大，并对相关数值进行区域标蓝。

4 结果

4.1 条件设定

在表 3 中，标蓝的区域为概率最大的取值，即可以被理解为出现的可能性最高的取值。我们将快艇流体动力学常数取值设定为相应概率所得进行判定的设定条件。即[1]=2, [2]=3,…如下表所示：

属性 / 取值	**[1] = 2**	**[2] = 3**	**[3] = 2**	**[4] = 1**	**[5] = 2**	**[6] = 3**
1	17.86	11.69	15.58	19.48	16.56	0
2	27.60	17.21	23.05	29.22	25.32	21.10
3	18.18	12.01	15.91	19.48	17.21	28.90

表 4 针对给定条件下的分别对残余阻力不同取值的概率计算。

Table 4. Probability calculations for different values of residual resistance under given conditions.

根据计算所得概率可以看出，在计算结果“1”中第 6 个属性的属性值为 0，这就导致计算结果出现概率为 0 的错误值。因此我们需要引入 Laplace 修正，另 λ=1. 即

$$P_\lambda(Y = c_k) = \frac{\sum_{i=1}^{N} I(y_i = c_k) + \lambda}{N + K\lambda}$$

在利用该修正后重新计算得属性 7 取值为 1 的前 6 属性概率如下表所示。

属性 / 取值	[1] = 2	[2] = 3	[3] = 2	[4] = 1	[5] = 2	[6] = 3
1	0.18	0.12	0.16	0.20	0.17	0.01
2	0.28	0.18	0.23	0.29	0.26	0.21
3	0.19	0.13	0.16	0.20	0.18	0.29

表 5 引入 Laplace 修正之后的概率计算

Table 5. Probability calculation after introducing Laplace correction.

因此，根据修正过后的新概率，我们可以对属性[7]的结果进行概率计算。

4.2 概率结果

最终，我们可以计算得到第 7 属性的三个取值分别对应概率如下所示：

1	2	3
1.20E-06	0.00018444	3.94E-05

最后将三个概率结果用图表示：

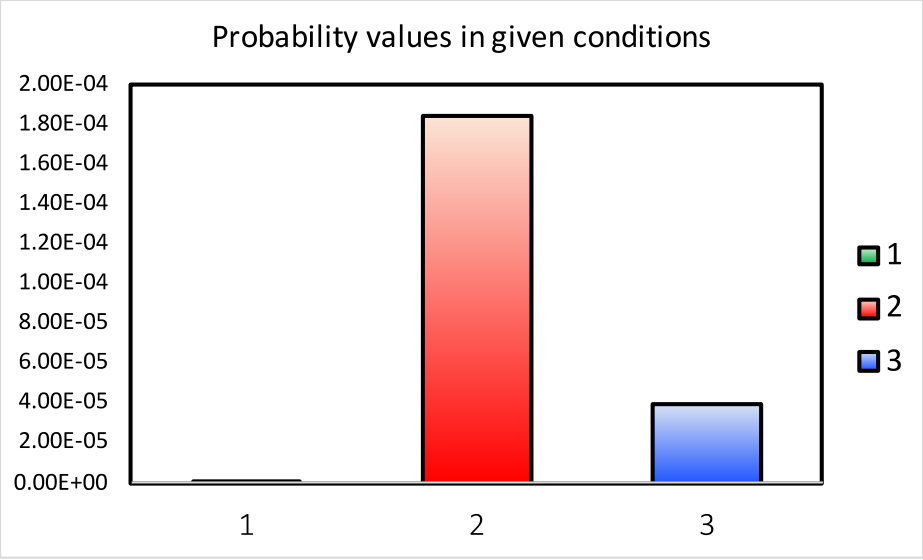

图 4 残余阻力取值为 1、2、3 的分别概率对应图.

Fig.4. Residual resistance values of 1, 2, and 3 respectively.

根据图 4 所得结果，我们可以判断出在给定的条件下，残余阻力为取值“2”的几率最大，如下所示。

$$P(Given\ Conditions|\mathbf{2}) > P(Given\ Conditions|\mathbf{3}) > P(Given\ Conditions|\mathbf{1})$$

5 结论

由关于游艇流体动力学 7 个参数给定特定数据库(表 1)，我们可以根据特定的分类取值方法(表 2)将数据库中 7 个属性分别对应的数据集由连续划分为三个取值(1,2,3)。快艇流体动力学数据分类前后的数据集具有相对不同的性质(图 3)。对于这些实验室中快艇流体动力学系数分级后的间断数据库相对更易分类分析。对于分类之后的数据库，我们可以计算前 6 个属性值的分别取值 1、2、3 的概率(表 3).由朴素贝叶斯计算结果，我们可以判断出快艇流体动力学这 6 个属性([1], [2],…[6])的分别那个值的取值概率最大。这 6 个属性的对应最高概率取值即是用于判断和计算属性 7 概率的设定条件。在设定条件下，利用朴素贝叶斯算法，我们可以计算设定条件下的属性 7 分别取值的概率(表 4)。但因为计算结果中属性 7 中出现取值为 1 对应属性[6]的取值为 1，故引入 Laplace 修正方法，得到重新计算所得概率结果(表 5).通过该表中对应概率结果，可以计算在设定条件下属性 7 取值 1、2、3 的概率。最终得到在设定条件下快艇残余阻力为等级“2”的概率最大。

6 总结

本文引用了 UCI 机器学习数据库中由代尔夫特理工大学海事和运输技术系船舶流体力学实验室提供的游艇水动力数据为分析数据库，作者同时对数据重新做了连续数据取间断值处理，并对于数据改变后进行作图分析(图 3)，以便于在 MATLAB 平台上进行朴素贝叶斯的运算。基于朴素贝叶斯算法，我们可以对给定的浮力中心纵向位置、棱镜系数、长度-位移比、束流比、长光束比、弗洛伊德数取值的条件下，通过概率判断游艇哪种残余阻力出现可能性最大。诚然，在实际快艇运行中，会有更多的因素影响残余阻力。但本文提供了一种用于判断游艇残余阻力取值的可行性方法。

感谢上海大学力学与工程科学学院胡国辉教授的讲授与讨论.

参考文献

[1] GE’s Marine Solutions One Neumann Way MD S156 Cincinnati, Ohio USA 45215 www.ge.com/marine GE Marine Gas Turbines for Frigates. March 2018

[2] 2019, Dua, Dheeru and Graff, Casey, 2017, {UCI} Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences http://archive.ics.uci.edu/ml

[3] P. Harrington(compile), Li.Y, Li.P, Qu.Y, Wang.B(translate)Machines Learning in Action 2013 (The People's Posts and Telecommunications Press) (in Chinese) [P. Harrington 编，李锐等 译 机器学习实战 2013 (人民邮电出版社)

[4] Zhou Zhihua. Machines Learning 2018 (Tsinghua University Press) (in Chinese) [周志华 机器学习 2018 (清华大学出版社)

[5] Hand, D.J.. (2009). Naive Bayes. The Top Ten Algorithms in Data Mining. 163-178. 10.1201/9781420089653.ch9.

[6] J. Gerritsma, R. Onnink, and A. Versluis. Geometry, resistance and stability of the delft systematic yacht hull series. In International Shipbuilding Progress, volume 28, pages 276â€“297, 1981.

[7] Yang, Feng-Jen. (2018). An Implementation of Naive Bayes Classifier. 301-306. 10.1109/CSCI46756.2018.00065.

Naïve Bayes classification calculation on Yacht Hydrodynamics Data Set

Zhai Hanfeng

(School of Mechanics and Engineering Science, Shanghai University, Shanghai 200444, China)

Abstract

Yacht are very common in nowaday's society. Yacht has a variety of complex parameter indicators during navigation. Relevant research analysis in the laboratory and recording the values of different attributes can help us understand how the speedboat dynamic parameters control the speedboat movement. The

naive Bayes algorithm is used to calculate the hydrodynamic data of the yacht provided by the Marine Hydrodynamics Laboratory of the Department of Maritime and Transportation Technology of Delft University of Technology. We can use the result of probability calculation to determine the residual resistance of the speedboat under specific settings. Because the data is calculated by the laboratory quantitative process, the results are considered continuous. In order to facilitate the naive Bayes operation, I used the discontinuous value classification method to classify the data into three levels [1, 2, 3] for probability calculations. According to the database, in the given conditions for based on the database probability calculation, we are able to analyze the which value of the residual resistence is more likely to occur (with the highest probability). The results of these analysis and calculations can help control the operating conditions of speedboats and better solve fluid dynamics related problems.

Keywords: Fluid mechanics; hydrodynamic parameters; naive Bayes; machine learning