北京市国产轻型轿车车载诊断系统（OBD）数据分析[1]

—基于数据挖掘的 OBD 参数研究

云南师范大学　　黄琼华、王敏、王璐

摘要

OBD（On-Board Diagnostics：车载诊断系统）是一种装置于车中用以监控车辆污染的系统，可于车辆的排放控制元件出现问题时，早期产生讯号以通知车主送店维修，避免问题车辆在不知情的情况下制造更多的污染和长期磨损对汽车造成伤害。OBD 技术还应用于汽车排污预测与防污措施建议等方面。由于 OBD 诊断数据专业性强且繁多，如何将指标简化易懂，有选择性的呈现给车主是个比较关键的问题。为找出这些指标，我们通过基于数据挖掘分析，对 OBD 数据集数据进行预处理，用聚类分析，主成分分析和因子分析将数据分类并筛选出比较重要的指标，然后用随机森林做预测模型，选取重要变量，最后用这些变量做回归，得出结论。希望能为车主和汽车制造商提供一些参考。

关键词：数据挖掘 聚类分析 主成分分析 因子分析 随机森林

[1] 注:该论文获得由中国统计教育学会举办的“ 2015 年（第四届）全国大学生统计建模大赛 ”大数据统计建模类本科生组一等奖。

一、问题的提出及研究思路

（一）问题的提出

OBD 是英文 On —Board Diagnostics 的缩写，即“车载诊断系统”。该系统可根据发动机的运行状况随时监控汽车尾气排放是否超标——一旦检测到汽车尾气排放超标，它会马上发出警示。当系统出现故障时，故障灯(MIL)或检查发动机(CheckEngine)警告灯亮，同时动力总成控制模块(PCM) 将故障信息存入存储器，通过一定程序可以将故障码从 PCM 中读出。根据故障码的提示，维修人员能迅速准确地确定故障的性质和部位，以缩短维修时间。

目前国内汽车技术还在发展，国外主要用于监控高排量汽车。OBD 系统可监测车辆多个系统和部件，包括发动机、三元催化转换器、颗粒捕集器、氧传感器、排放控制系统、燃油系统、废气再循环系统(EGR)等。OBD 技术的引入和扩展，是对汽车产业链的一个考验和提高。OBD 技术能通过所记录的综合参数和即时参数有效监测汽车尾气排放。2005 年 4 月 5 日，国家环境保护总局公告(2005)14 号颁布《轻型汽车污染物排放限值及测量方法（中国 III、IV 阶段）》GB 18352.3–2005，正式明确了我国对 OBD 系统的技术要求。但由于 OBD 诊断数据专业性强且指标繁多，车主面对复杂的数据会感到抓不住要领从而放弃使用 OBD，这个结果显然是我们不愿意看到的，如何将指标简化易懂，有选择性的呈现给车主是个比较关键的问题。

特别要强调的是，我国面临着日益严峻的环境污染问题，中东部大部地区雾霾频发，而城市雾霾的首要形成因素是汽车尾气，其次，机动车的尾气是雾霾颗粒组成的最主要的成分，最新的数据显示，北京雾霾颗粒中机动车尾气占 22.2%，燃煤占 16.7%，扬尘占 16.3%，工业占 15.7%。但随着汽车技术进步以及油品质量的上升，环境管理者发现机动车尾气对雾霾天气形成并不起决定性作用，但作为一些汽车拥有量较大的城市，管理者依旧需要控制机动车排放标准，避免雾霾天气的形成。基于此，OBD 技术的推广非常有必要，同时 OBD 诊断的数据如何能让车主一目了然并能发现汽车所存在的故障，即时送去检修是我们的研究问题，我们希望通过数据挖掘技术，能找出人们需要关注的主要指标。

（二）研究思路

首先，我们用数据挖掘技术将数据聚类，看看能否分为 9 类（因为数据中来自 9 类车型），先用分层聚类自动寻求分为多少类，看看各类中这 9 类车分布如何，没有集中的迹象的车子特征不明显。然后预测所有的，看看哪些车会分到其他车型中，错分的越多说明这些车子相似，不易区分。可以看出车辆的相似性随着变量类型不同而不同。其次，我们对数据集做主成分分析和因子分析，将数据指标降维，看看哪些指标说明一些共同问题。最后，用随机森林提取重要指标预测模型，用重要指标建立多元线性回归模型。

二、建模准备

（一）数据来源

http://datatang.com/五月限免数据—OBD 数据集。

OBD 数据集包含三个数据表：OBD_INFO_HISTORY 综合参数、OBD_INFO_IMMEDIATE_HISTO-RY 即时参数、OBD_TERMINAL 带 VIN 码的车辆数据。综合参数包含油耗、长时燃油修正、短时燃油修正、燃油压力计量、点火正时、绝对节气门、歧管真空度油轨压力、进气管压力、进气温度、空气流量、电瓶电压(行车阶段)等以及服务器响应信息等数据；即时参数包括速度、引擎以及服务器响应消息等数据；带 VIN 码的车辆数据包含车辆 ID、当前总里程、当前城市编码、当前城市名称、车辆类型、VIN 码等数据项。

（二）数据分析与预处理

综合参数表中含有 12544 行观测值和 75 个参数，即时参数表中含有 184927 行观测值和 14 个参数，其中有几个参数用“＃”连接，我们用 Excel 进行了拆分。通过查阅相关书籍和文档，我们发现汽车油耗、节气门开度、冷却液温度、氧传感器的工作状态、进气歧管的压力（绝对压力）、点火正时、速度、引擎等指标可能是反映汽车故障的重要指标，同时发现它们之间可能存在线性相关，而服务器响应消息、VIN 码和 EGR 开度、入库时间可能无影响故删除。长、短时燃油修正、氧传感器监测指标包含四个值，我们用 Excel 进行了拆分。OBD 数据集有 10 种车型的信息，通过查询，发现其中一种车型（起亚轻型客车）与其他 9 种车车型差别较大，不好一起分析，所以删除了这辆车的数据。因为数据集观测值较多，剔除个别不完整数据后的数据集不含缺失值。在综合参数数据表中，我们对内部 ID 重新编码如下：

表 1

内部 ID	重新编码	车辆类型
352016850081381	1	长安 SC7134CB5 轿车
352016818128522	2	奇瑞 SQR7160ES 轿车
352016889385480	3	奇瑞 SQR7151A217 轿车
352016864499108	4	颐达 DFL7160AB 轿车
352016857330054	5	长安 SC7134C 轿车
352016802932327	6	速腾 FV7146TATG 轿车

352016857330088	7	长安 SC7139A4B 轿车
352016840869622	8	长安 SC7150G 轿车
352016859398034	9	长安 SC7150G 轿车

对于即时参数数据表，以示区分，没有进行编码。

（三）模型假设

1、在做机器学习和多元线性回归时，假设平均油耗为我们所选取的因变量。
2、进行数据预处理所删除的变量都是不重要的变量。
3、多元回归拟合模型满足 Guass-Markov 假设条件。

（四）变量选择

汽车污染的排放量及其许多自身的性能都与耗油量有关，因此把油耗量作为因变量；我们通过随机森林，作出变量重要性图，可以看出与油耗量相关的主要变量；另外，通过查阅相关资料，我们了解到与汽车油耗量相关的因素主要有：进气温度、空气流量、冷却液水温、氧化器温度等等，因此可以把这些变量作为自变量，具体的回归及其检验将在下面的部分详细介绍。

三、模型构建

多元统计分析是多变量的统计分析方法，包含了丰富的理论成果和众多的应用方法。它主要包括回归分析、方差分析、判别分析、聚类分析、主成分分析、因子分析和典型相关分析等。对于 OBD 数据集，我们主要采用了回归分析、聚类分析和主成分分析等方法。另外，还加入了现代机器学习方法，目的是从不同的角度对 OBD 数据集进行统计建模和分析，结论更完善。

（一）聚类分析

聚类分析是研究“物以类聚”的一种方法。聚类分析是应用最广泛的一种分类技术它把性质相近的个体归为一类，使得同一类中的个体具有高度的同质性，不同类之间的个体具有高度的异质性。聚类分析的职能是建立一种分类方法，它是将一批样品或变量，按照它们在性质上的相似程度进行分类。通常我们用距离来度量样品之间的相似程度，用相似系数来度量变量之间的相似程度。

（二）主成分分析

也称主分量分析，是一种将多个指标化为少数几个综合指标的统计分析方法。在经济问题研究中，为了全面、系统地分析问题，我们必须考虑众多对某经济过程有影响的因素，这些因素在统计学中被称为指标，也称为变量，每个指标都在不同程度上反映了所研究问题的某些信息，但是指标之间彼此有一定的相关性，因而所得的统计数据在一定程度上反映的信息有重叠。主成分分析可将相关的指

标化成少量不相关的指标，避免了信息重复带来的虚假性。此外，主成分分析能用较少的变量反应更多的问题，减少计算量的同时简化了问题。

（三）多元线性回归

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

X_1，X_2，X_3，X_4，X_5，X_6分别为下面我们建立的模型的解释变量。

（四）机器学习方法

最初的数据挖掘分类应用大多都是在这些方法及基于内存基础上所构造的算法。目前数据挖掘方法都要求具有基于外存以处理大规模数据集合能力且具有可扩展能力。下面对几种主要的分类方法做个简要介绍。

1.决策树

决策树归纳是经典的分类算法。它采用自顶向下递归的各个击破方式构造决策树。树的每一个结点上使用信息增益度量选择测试属性。可以从生成的决策树中提取规则。

2.SVM 法

SVM 法即支持向量机(Support Vector Machine)法，由 Vapnik 等人于 1995 年提出，具有相对优良的性能指标。该方法是建立在统计学习理论基础上的机器学习方法。通过学习算法， SVM 可以自动寻找出那些对分类有较好区分能力的支持向量，由此构造出的分类器可以最大化类与类的间隔，因而有较好的适应能力和较高的分准率。该方法只需要由各类域的边界样本的类别来决定最后的分类结果。支持向量机算法的目的在于寻找一个超平面 H(d),该超平面可以将训练集中的数据分开，且与类域边界的沿垂直于该超平面方向的距离最大，故 SVM 法亦被称为最大边缘(maximum margin)算法。待分样本集中的大部分样本不是支持向量，移去或者减少这些样本对分类结果没有影响，SVM 法对小样本情况下的自动分类有着较好的分类结果。

3.boosting 回归

训练集中一共有 n 个点，我们可以为里面的每一个点赋上一个权重 $W_i(0 \le i \le n)$，表示这个点的重要程度，通过依次训练模型的过程，我们对点的权重进行修正，如果分类正确了，权重降低，如果分类错了，则权重提高，初始的时候，权重都是一样的。可以想象得到，程序越往后执行，训练出的模型就越会在意那些容易分错（权重高）的点。当全部的程序执行完后，会得到 M 个模型，分别对应$Y_1(X) \cdots Y_M(X)$，通过加权的方式组合成一个最终的模型 $Y_M(X)$。

我们觉得 Boosting 更像是一个人学习的过程，开始学一样东西的时候，会去做一些习题，但是常常连一些简单的题目都会弄错，但是越到后面，简单的题目已经难不倒他了，就会去做更复杂的题目，等到他做了很多的题目后，不管是难题还是简单的题都可以解决掉了。

4.随机森林回归

随机森林（random forest）模型是由 Breiman 和 Cutler 在 2001 年提出的一种基于分类树的算法。它通过对大量分类树的汇总提高了模型的预测精度，是取代神经网络等传统机器学习方法的新的模型。随机森林的运算速度很快，在处理大数据时表现优异。随机森林不需要顾虑一般回归分析面临的多元共线性的问题，不用做变量选择。现有的随机森林软件包给出了所有变量的重要性。另外，随机森林便于计算变量的非线性作用，而且可以体现变量间的交互作用（interaction）。它对离群值也不敏感。

四、数据挖掘技术

（一）理论准备

1.K 均值聚类法

系统聚类法的每一步都要计算"类间距离",计算量比较大,特别是当样本量比较大的时候,系统聚类需要占很大的内存空间,计算也比较费时间,为了改进这个不足,Mac Queen 提出了一种动态快速聚类方法------K 均值聚类法,其基本思想是:根据给定的参数 K,先把 n 个对象粗略地分为 K 类,然后按照某种最优原则(通常表示为一个准则函数)修改不合理的分类,直到准则函数收敛为止,这样就得到了最终的分类结果。

2.主成分分析

在用统计方法研究多变量问题时，变量太多会增加计算量和增加分析问题的复杂性，我们希望在研究问题的过程中，涉及的标量较少，代表的信息较多。于是产生了主成分分析。在实际操作中将原来 p 个指标作线性组合，作为新的综合指标。其中选取的第一个线性组合，即第一个综合指标，记作 F_1，$Var(F_1)$ 越大，表示 F_1 包含的信息越多，并称 F_1 为第一主成分。第一主成分不足以代表原来 p 个指标的信息时，再考虑选取 F_2 即选第二个线性组合。为了有效地避免信息重叠，

F_1已有的信息就不需要再出现在F_2中，用数学语言表达即$Cov(F_1,F_2)=0$，此时称F_2为第二主成分，依此类推可以构造出第三、第四，……，第p个主成分。

主成分的定义：

$$\begin{cases} F_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p \\ F_2 = a_{12}X_1 + a_{22}X_2 + \cdots + a_{p2}X_p \\ \vdots \\ F_p = a_{1m}X_1 + a_{2m}X_2 + \cdots + a_{pm}X_p \end{cases}$$

上述方程组要求：

$a_{1i}^2 + a_{2i}^2 + \ldots\ldots + a_{pi}^2 = \lambda_i (i = 1, \ldots\ldots, m)$；

$A'A = I_m$，（$A = \left(a_{ij}\right)_{p\times q} = (\alpha_1, \alpha_2, \cdots, \alpha_m)$ A为正交矩阵）；

$Cov(F_i, F_j) = \lambda_i \delta_{ij}$，$\delta_{ij} = \begin{cases} 0, i \neq j \\ 1, i = j \end{cases}$；

$a_{t1} + a_{t2} + \ldots + a_{tm} = 1 \quad (t = 1, \ldots, p)$。

其中：$\mathrm{A} = (a_{ij})_{p\times q} = \left(\alpha_1, \alpha_2, \cdots, \alpha_m\right); (a_{1i}, a_{2i}, \cdots, a_{pi})$（其中$i = 1, \cdots, m$）为 X 的协差阵Σ的特征值对应的特征向量；$X_1, X_2, \cdots, X_p$是原始变量经过标准化处理的值（因为在实际应用中，往往存在指标的量纲不同，所以在计算之前先消除量纲的影响，将原始数据标准化）；$R\alpha_i = \lambda_i \alpha_i$，其中$R$ 为相关系数矩阵，λ_i、α_i是相应的特征值和单位特征向量，有$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$。

如上式所述：这样就用m个综合指标对原来的p $(m < p)$个指标的数据进行提炼和阐述，提取综合贡献率达到70%或80%的前几个主成分对总体进行排名和分析。

采用主成分分析建立模型的一般步骤如下：第一步，设样本p个指标为$X_1, X_2, \ldots, X_p$，共n个，为了消除量纲的影响以及各指标在数量级上的差别，将原始数据采用（1） 进行标准化处理

$$x_{ij}^* = (x_{ij} - E_j) \ / \mathrm{S}_j$$

其中，x_{ij}^*是x_{ij}的标准化数据，E_j和S_j分别是第j个指标的样本均值和样本标准差；

第二步，建立标准化数据的相关系数矩阵$R = (r_{ij})_{p\times p}$，r_{ij}是X_i^*与X_j^*的相关系

数；

第三步，求出相关矩阵 R 的特征值（λ_1，$\lambda_2,\cdots,\lambda_p$），及对应的特征向量 $\mu_1,\mu_2,\cdots,\mu_p$，其中 $u_i=(u_{i1},u_{i2},\ldots,u_{ip})$（其中 $i=1,2,\ldots,p$），于是得到 p 个主成分：

$$F_i=u_{i1}X_1*+u_{i2}X_2^*+\ldots+u_{ip}X_p^*$$

其中 F_i 是第 i 个主成分（$i=1,2,\ldots,p$）

第 i 主成分 F_i 的特征值（λ_i）即为该主成分的方差，方差越大，对总变差的贡献也越大，其贡献率为 λ_i/p，其中（贡献率 λ_i/p）反映了第 i 主成分综合原始变量信息的百分比；

第四步，确定主成分。

（1）.根据特征值大小确定，一般取大于 1 的特征值。

（2）.根据主成分的累计方差贡献率来确定，即选取使得主成分的累计贡献率大于等于 70%或 80%的最小整数 m，就确定了前 m 个主成分。

3.现代分类和回归：机器学习算法的基本思想

有些回归和分类模型是可以写成公式的。但是另外一些回归和分类的方法是体现在算法之中，其具体形式是计算机程序，这些方法广泛用于机器学习或数据挖掘之中。算法模型适用的范围比经典的统计模型更加广泛。由于现在经典模型也要经过计算机软件实现，因此，总的来说，算法模型包含了经典模型，只不过算法模型和经典模型的发展过程及思维方式有很大的不同。算法建模主要发展于最近二十年，它得益于不断进步的计算机技术。如果说起源于前计算机时代的经典统计目前大大受益于计算机的发展，那么没有计算机，就不可能产生算法建模。

在处理巨大的数据集上，在对付被称为维数诅咒的巨大变量数目时，在无法假定任何分布背景的情况下，在面对众多竞争模型方面，算法建模较经典建模有着广泛的应用及理论前景。

4. 多元线性回归

（1）.模型定义

多元线性模型通常用来描述变量 y 与 x 之间的随机线性关系,即

$$y=\beta_0+\beta_1x_1+\cdots+\beta_kx_k+\varepsilon$$

式中,X 是非随机变量;Y 是随机的因变量;β_0 是常数项;β_i 是回归系数; ε 是随机误差.

如果对 y 和 x 进行了 n 次观测,得到 n 组观测值 $y_i,x_{1i},x_{2i},\ldots,x_{ki}$ (i=1,2,3…n),他们满足以下关系式

$$y_i=\beta_0+\beta_1x_{1i}+\cdots+\beta_kx_{ki}+\varepsilon_i$$

引入矩阵记号,记为

$$Y=\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad X=\begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & x_{12} & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix} \quad \beta=\begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix} \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

则模型可以写成如下矩阵形式

$$Y=X\beta+\varepsilon$$

式中,Y 是 nx1 观测向量;X 是 nx(k+1)已知设计矩阵; ε 是 nx1 随机误差向量;β是(k+1)x1 未知参数向量。

如果模型满足 E(ε)=0;Var(ε)=σ^2I ;x_1， x_2， …， x_k (k=1,2..)互不相关,则称模型为普通线性回归模型。

（二）聚类分析

1.基于主成分分析的分层聚类

在实践中,直接聚类的结果往往不太好,原因是有噪声干扰,但是如果先进行主成分分析,再进行聚类,结果可能会好得多。

我们先对综合参数进行分层聚类，得到如下结果（图 1 到图 3），从图中可以清晰的看出将数据分为 8 类，5 号（长安 SC7134C 轿车）和 8 号（长安 SC7150G 轿车）数据分为同一类，而 9 号（长安 SC7150G 轿车）车没有集中迹象，为单独的一类。

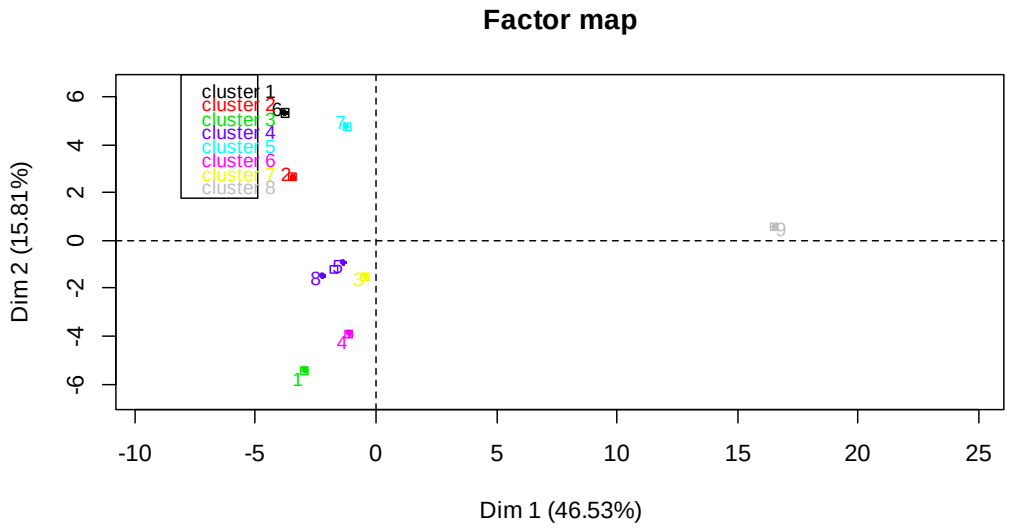


图 1 综合参数分层聚类

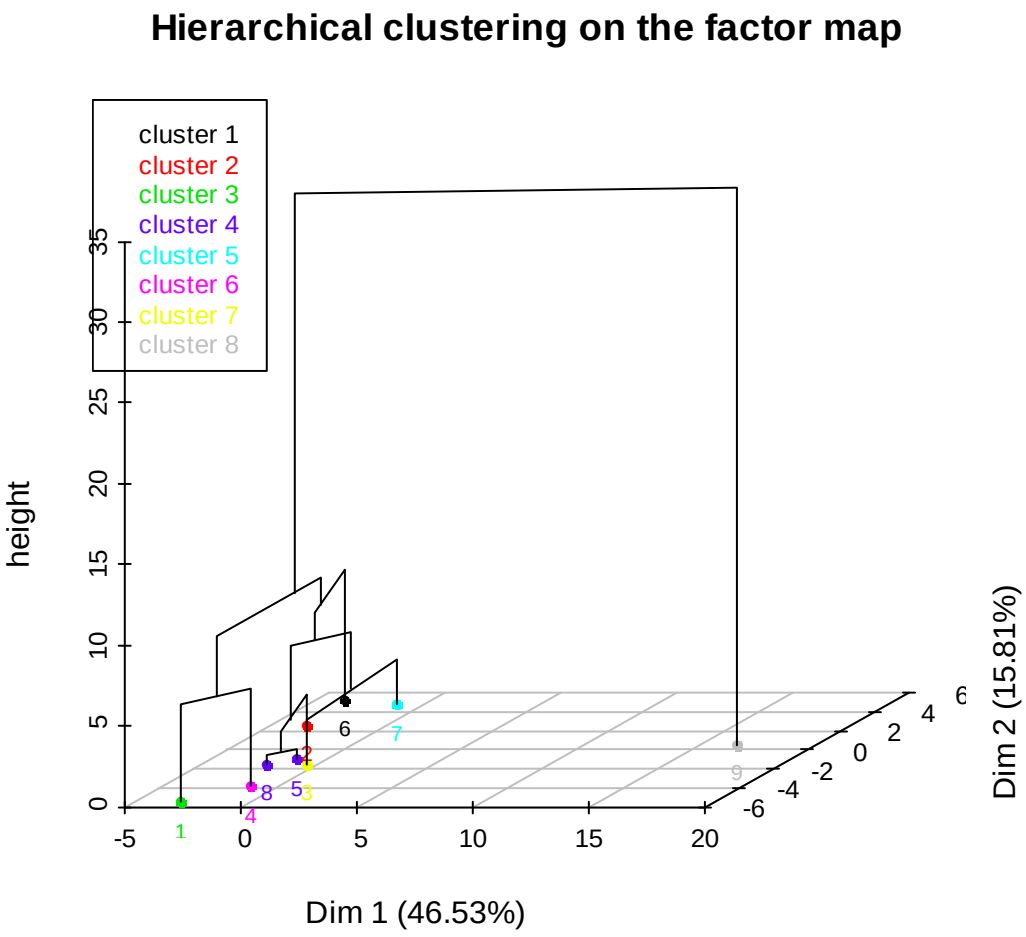


图 2 综合参数分层聚类

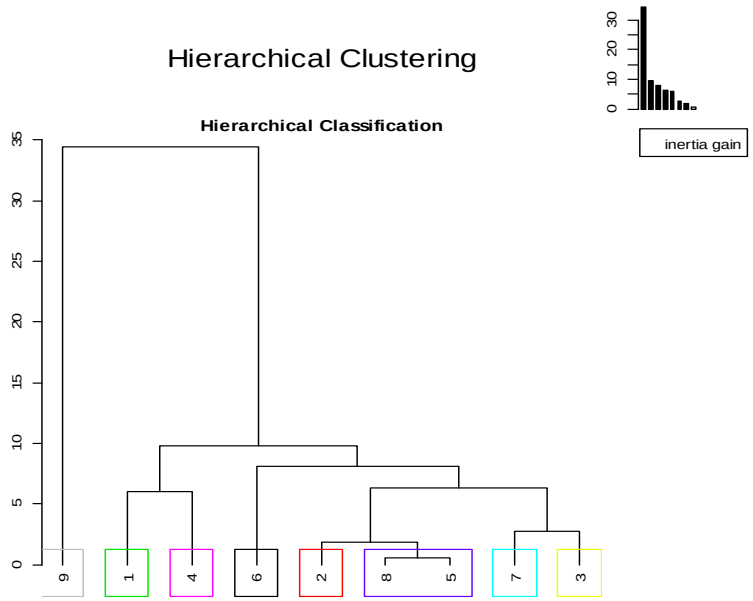


图 3 综合参数分层聚类

接着对即时参数进行分层聚类，如下三个图（即时参数分层聚类图 4 到图 6），与综合参数类似，也是分为 8 类，而内部 ID 为 352016859398034 的车，也就是综合参数中编码为 9 号（长安 SC7150G 轿车）的车离群，推测为特征明显，单独分为一类。

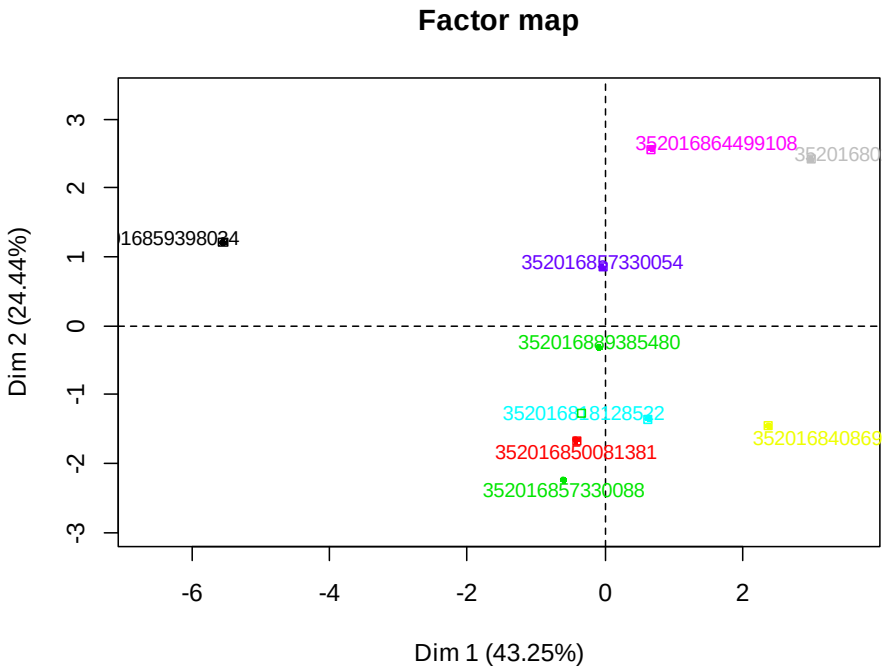


图 4 即时参数分层聚类

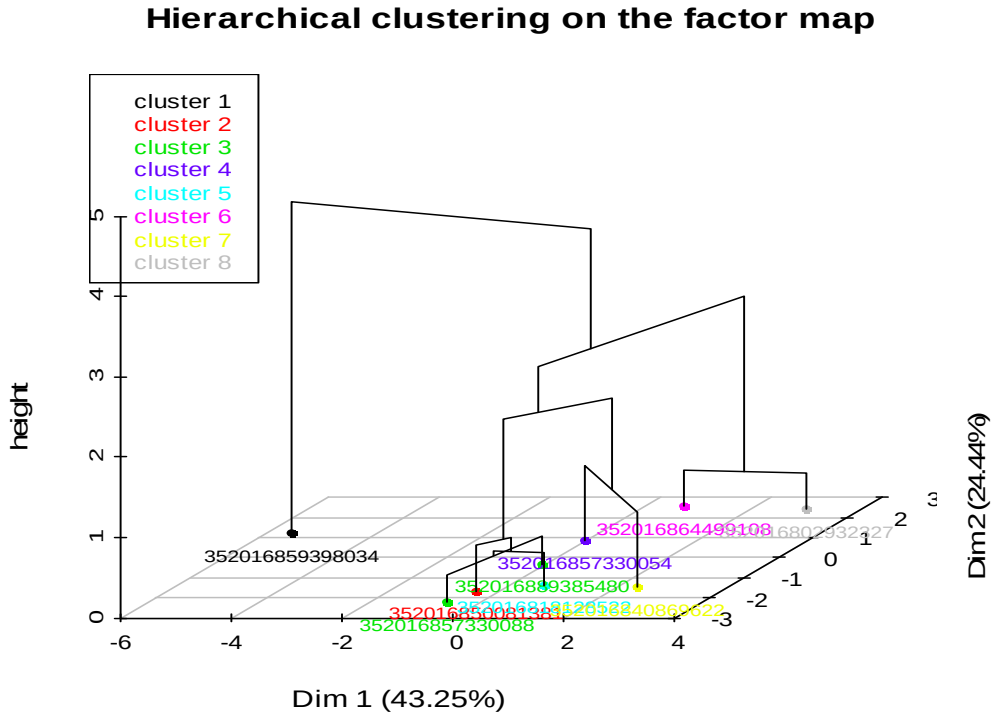


图 5 即时参数分层聚类

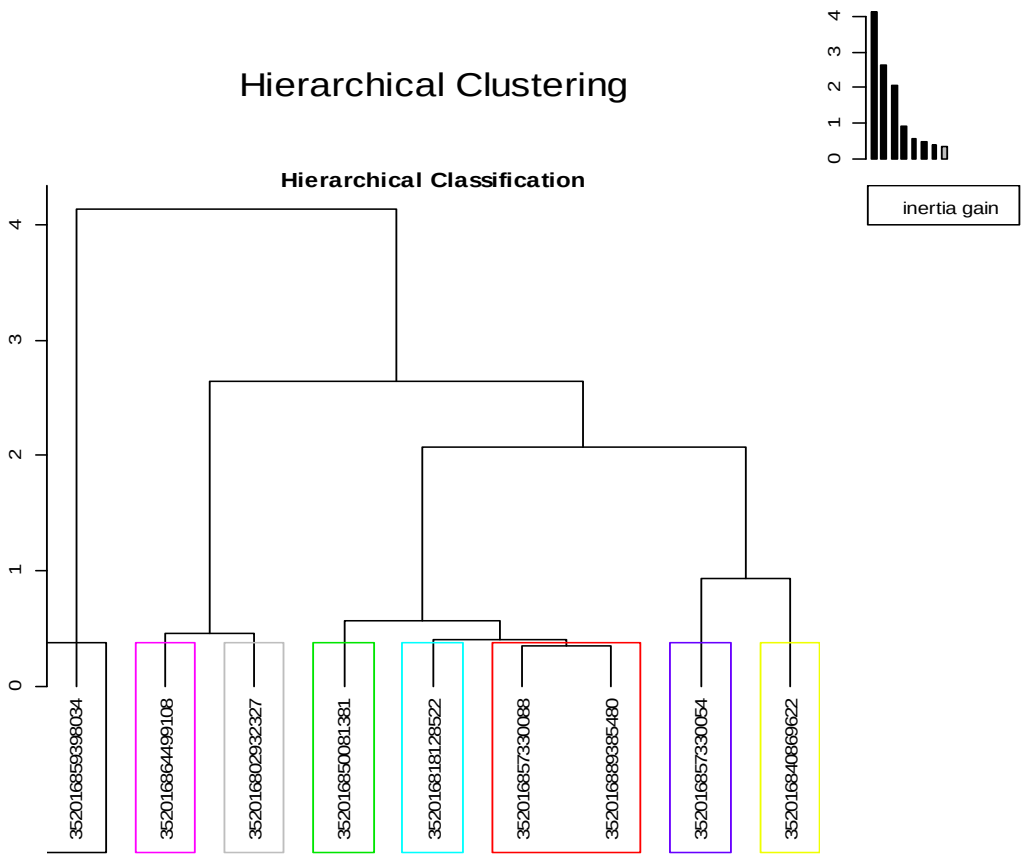


图 6 即时参数的分层聚类

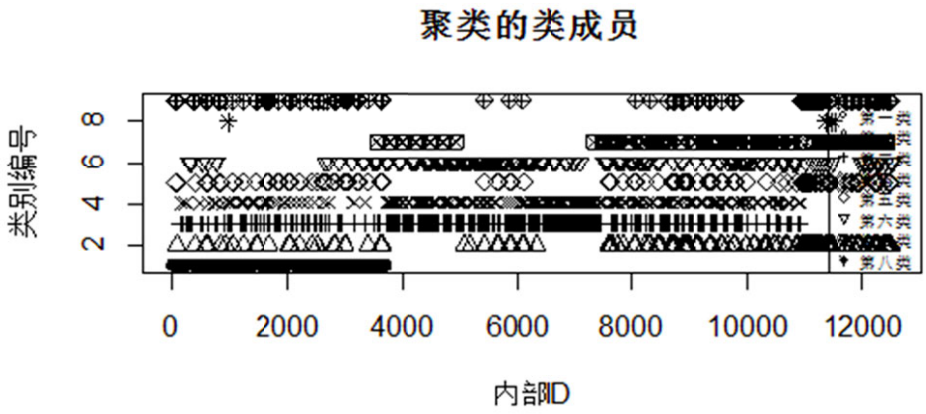
2.K-Means 聚类

综合参数的 K-Means 将数据聚为 8 类，由于样本量太多，图（图 7）不是很清晰，我们用 plot 函数和 lines 函数绘了 8 类的成员情况以及各类变量均值的变化折线图，直观展现了 8 类的结构特征。

评价类间的差异性和相似性。

```
CluR$betweenss/CluR$totss*100
# 99.87382#
```

说明：因类间解释的离差平方和占总平方和的 99.87382% 总体聚类效果较好，聚类结果可以接受。



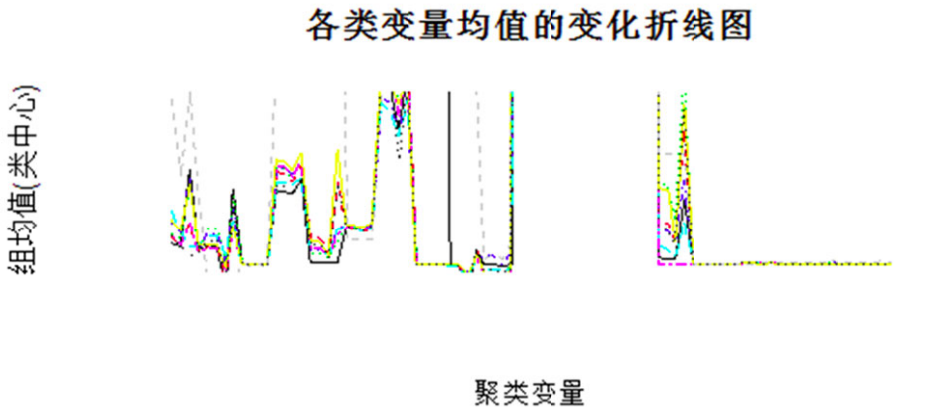


图 7 综合参数的 K-Means 聚类

即时参数的 K-Means 将数据聚为 8 类，由于样本量太多，图（图 8）不是很清晰，我们用 plot 函数和 lines 函数绘了 8 类的成员情况以及各类变量均值的变化折线图，直观展现了 8 类的结构特征。

评价类间的差异性和相似性。

```
CluR$betweenss/CluR$totss*100
#82.41624#
```

说明：因类间解释的离差平方和占总平方和的 82.41624%，总体聚类效果好，聚类结果可以接受。

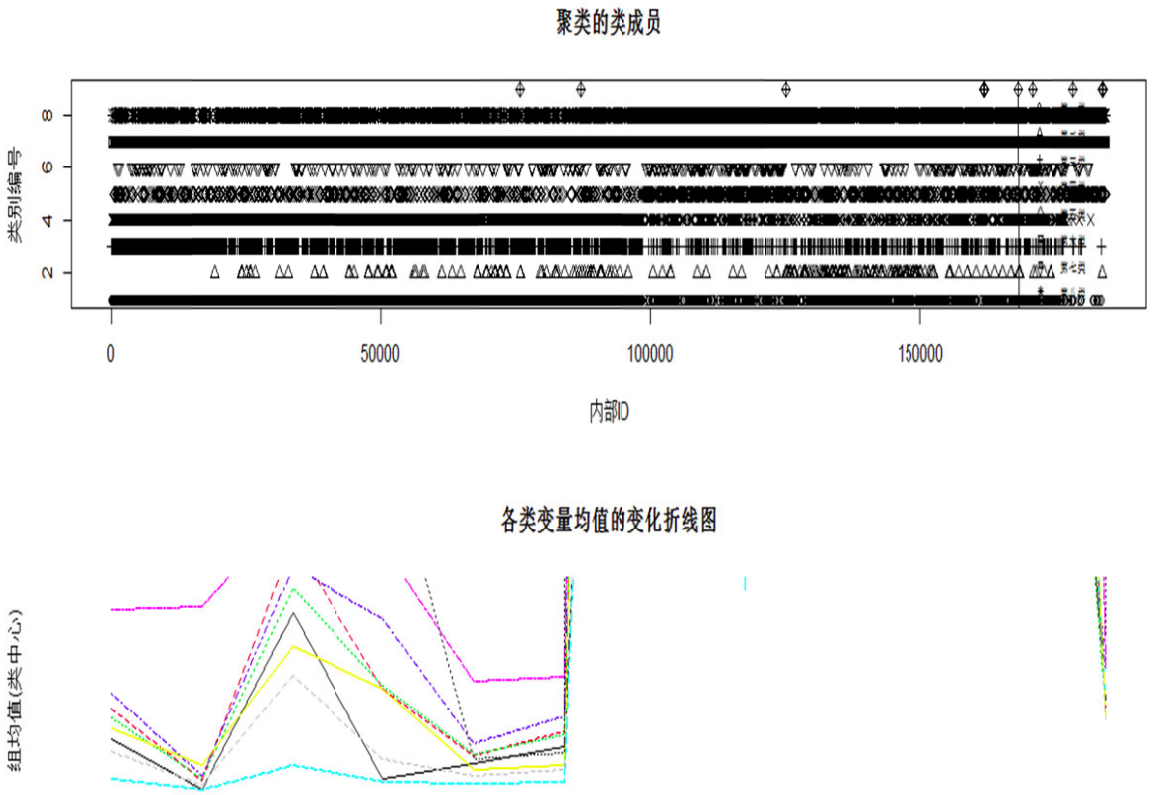


图 8 即时参数的 K-Means 聚类

聚类的类成员图的图例中可以看到将数据分为八类，比较基于主成分分析的聚类与 K-Means 聚类，结果大体相同，都是分为 8 类，分层聚类结果比较直观清晰，K-Means 可能由于样本点较多，图不是很清晰。

（三）主成分分析

1.本研究中是对以下 13 个变量做主成分分析:”平均速度","最小速度","最大速度","当前速度","当前里程.前一次打火时间到当前里程数.","总里程","当前引擎转速","当前引擎温度","前一次时间段至此引擎最大转速","前一次时间段至此引擎最小转速","前一次时间段至此引擎平均转速","引擎负荷”可以得到相应的特征根碎石图(图 9)和特征根累积贡献图(图 1),可见直到第五主成分时,累积贡献率才达到 75%左右,说明用主成分分析降维的效果并不是特别突出。

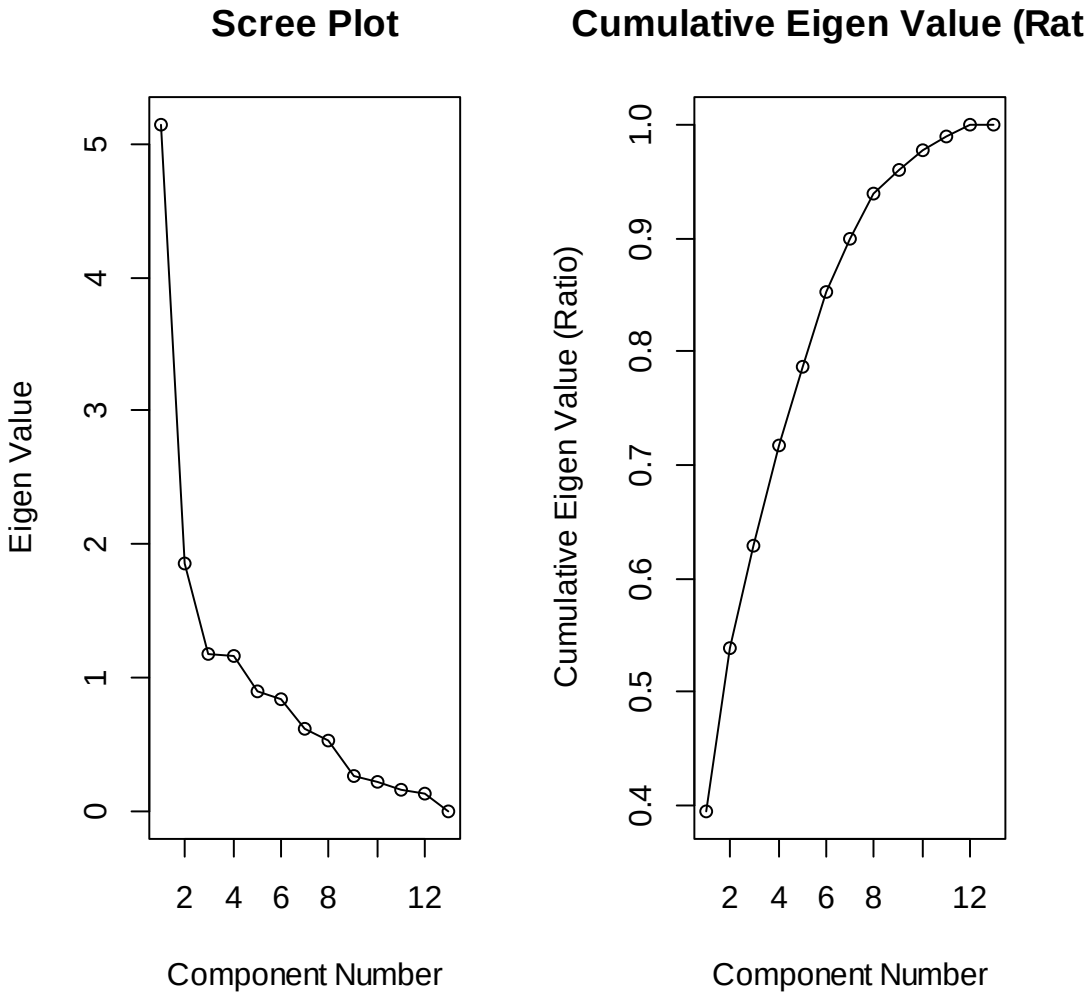


图 9 即时参数主成分分析的崖底碎石图(贡献率，左)和图 10 累计贡献率（右）

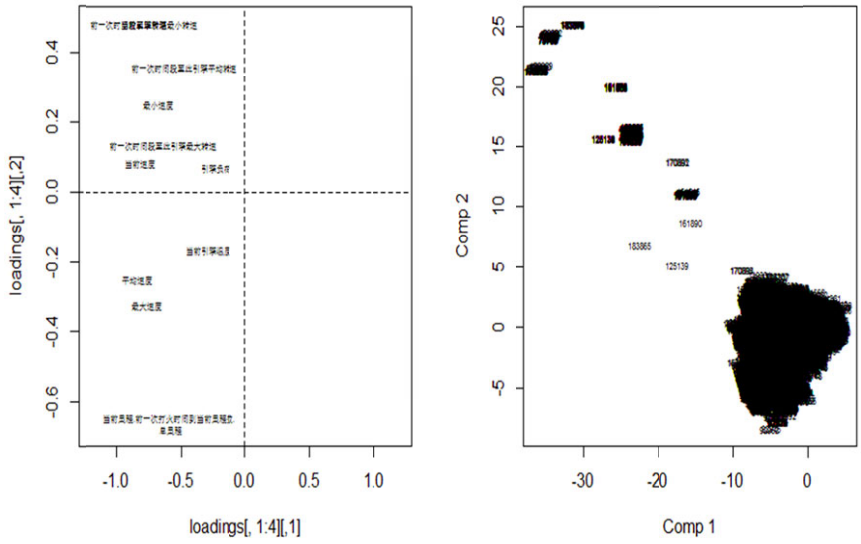


图 11 即时参数主成分分析的载荷图

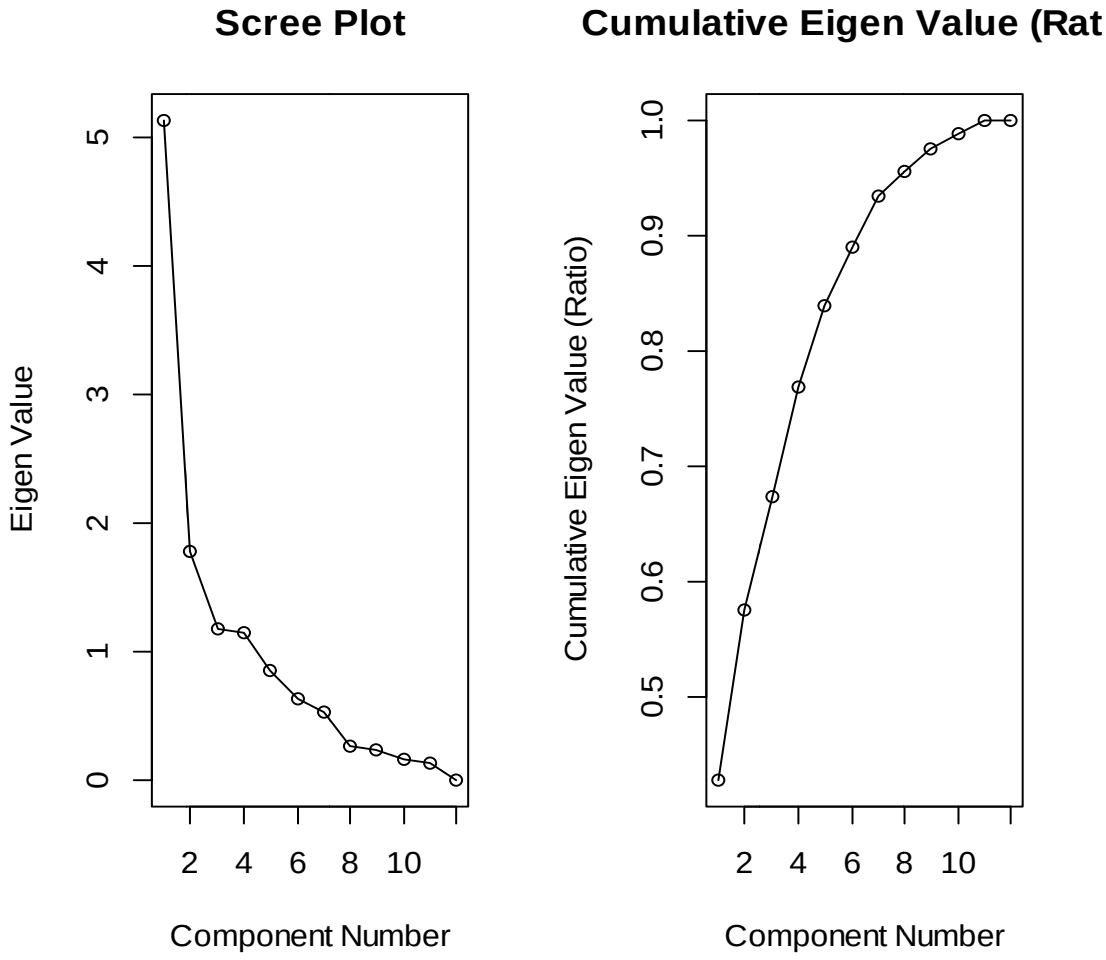


图 12 综合参数主成分分析的崖底碎石图(贡献率，左)和图 13 累计贡献率图（右）

```
> u
              [,1]         [,2]          [,3]          [,4]         [,5]
 [1,] -0.0749873998  0.109433984  0.0366250676  0.0482596421  0.166377538
 [2,] -0.0540597743  0.110541487  0.0604443193  0.0629696399  0.142382602
 [3,] -0.0197754486 -0.022626561  0.0778652724  0.0712049148 -0.054058264
 [4,] -0.0897518152  0.075313999  0.0046712113  0.0569757991  0.078094010
 [5,]  0.0277477703  0.002619782  0.1457425555  0.0608546231 -0.060127438
 [6,]  0.0381906938  0.008821278  0.1917634492  0.0704611018 -0.083818913
 [7,]  0.0107190053  0.132032439  0.1018970982  0.1025893307  0.044244197
 [8,]  0.0444323697 -0.158567749  0.1631244188 -0.0277249213 -0.145343617
 [9,] -0.0400249742  0.109436847 -0.1131842264  0.2147936421  0.064524202
[10,]  0.0083029584  0.020048244 -0.0050875114  0.0550299259  0.044765180
[11,] -0.0453537443  0.121189076 -0.1255191959  0.2344614475  0.069718240
[12,]  0.0113079887  0.028030994 -0.0014831450  0.0148898060  0.038645035
[13,] -0.0662249937  0.209473565 -0.0066523391 -0.1857653294  0.086664630
[14,] -0.0676347447  0.207863394 -0.0042040545 -0.1865998047  0.089310689
[15,] -0.0755907297  0.184956240  0.0134997738 -0.1582027561  0.131745872
[16,] -0.0645442132  0.186328056 -0.0006971596 -0.2029892035  0.015470151
[17,] -0.1962312125  0.056264908  0.0055504980  0.0566798936 -0.094752998
[18,] -0.2168545597  0.060478285  0.0082811128  0.0631300575 -0.103476178
[19,] -0.2407374424  0.023841842  0.0107781637  0.0165390386 -0.045620822
[20,] -0.1135951736  0.132336093 -0.0125390808  0.0924815642 -0.139998406
[21,]  0.0145151005  0.121893826 -0.0751727420  0.2152448946 -0.342254341
[22,]  0.0144026548  0.121759920 -0.0758821554  0.2158845266 -0.344047769
[23,]  0.0069041309  0.143380963 -0.0659058082  0.1651703285 -0.320098778
[24,]  0.0166791430  0.099303407 -0.0759350556  0.2175948667 -0.352561975
[25,]  0.0048743115  0.057298149 -0.0659224542 -0.3299830046 -0.182570310
[26,]  0.0032564228  0.058561983 -0.0507597050 -0.3276236201 -0.194625896
[27,] -0.0085551793  0.107351973  0.0164637062 -0.2445451931 -0.059740038
[28,]  0.0044152974  0.037661511 -0.0483298143 -0.3392961292 -0.181656522
[29,] -0.1184259327 -0.229946926 -0.0482875717  0.0125068692 -0.076325862
[30,] -0.1184217700 -0.229949252 -0.0482885626  0.0125077676 -0.076325987
[31,] -0.1184092607 -0.229956239 -0.0482915233  0.0125104730 -0.076326355
[32,] -0.1184369995 -0.229942794 -0.0482852877  0.0125064691 -0.076323645
[33,] -0.1788045143  0.015516930  0.0116199362 -0.0102607947 -0.019524844
[34,] -0.2199566779  0.017713014  0.0170766855 -0.0103222170 -0.017000480
```

图 14 综合参数的主成分得分图（主成分太多无法表示出来）

（四）现代分类和回归：机器学习方法

在 OBD 综合参数里，我们把平均油耗看作因变量，其他均看作自变量。

1.决策树回归：回归树

决策树所能处理的问题非常广泛，直观易懂，容易解释，这是传统统计所不可比拟的。用 R 画出的决策树如下所示（图 15）：

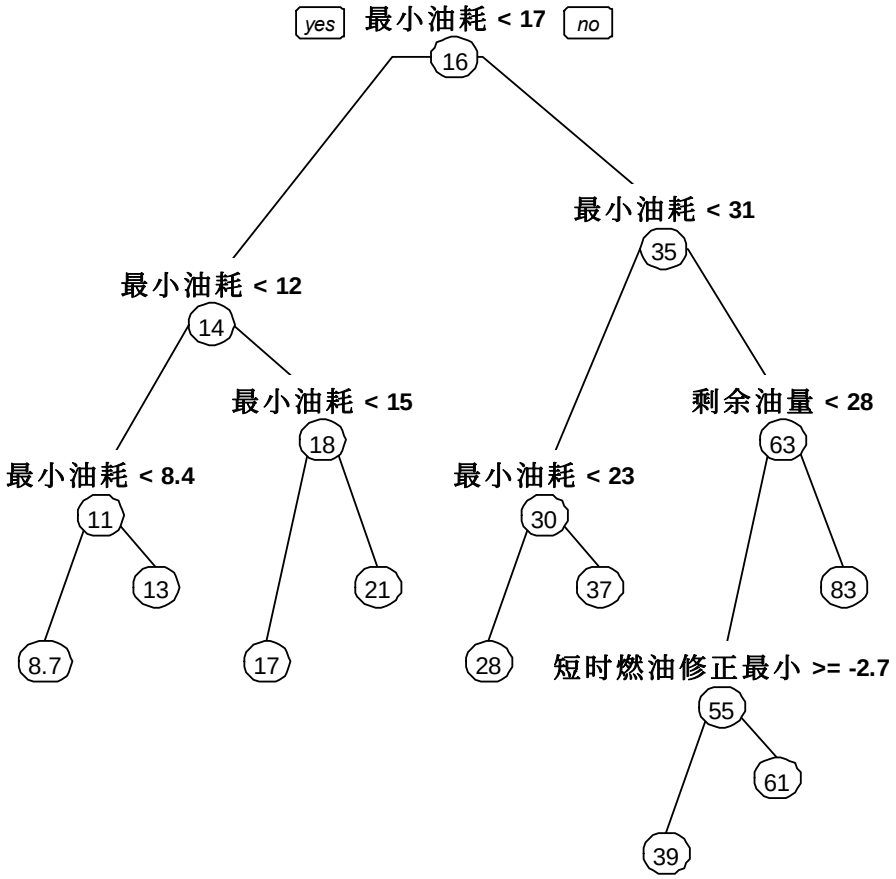


图 15 决策树

决策树就像一棵从根长出来的树（这里是倒着长的，也有横着长的）。最上面一个叫根节点，占据那里的变量为最小油耗。然后根据最小油耗是否大于等于 17 做出下一步决策，如果“是”（yes）则走向左边，“不是”（no）则走向右边；当走向左边时（最小油耗小于 17 的）数据就少一点。从根节点往右走，就进入另一个节点。到某个节点，决策树的这个分支结束了，这个节点称为叶节点或终节点。这个决策树有 9 个叶节点。有些值可能被误分到其他类别中，计算得误判率为 0.204176。这个决策树仅用了 85 个自变量中的 3 个。

2.boosting 回归

决策树一开始可能较弱（即出错率较高），然后，随着迭代的进行，不断地通过自助法加权再抽样，根据产生新样本来改进分类器，每一次迭代时都针对前一个分类器对某些观测值的误分缺陷加以修正，通常的做法是在（放回）抽取样本时对那些误分的观测值增加权重（相当于对正确分类的减少权重），这样在新的样本中就可能有更多的前一次分错的观测值，再形成一个新的分类器进入下一轮迭代，作为结果，这些观测值在训练模型时就有了更大的代表性，增加了对这类观测值的正确划分的可能性。由 R 的运行结果可知，boosting 回归的误判率为：0.1352。

3.支持向量机（SVM）

SVM 不是基于决策树的组合方法，它虽然是基于数学模型的，但此方法结合了计算机的算法。由 SVM 发展出来的回归方法也叫支持向量回归。对于线性不可分问题，可以做变换，使之成为线性可分问题。由于线性可分问题通过 Lagrange 乘子法的解仅仅涉及内积（对偶性质），线性不可分问题就变成简单地用某个核函数来代替单独变换的内积。回归用的 SVR 仅仅是把 SVM 的思想推广。该方法之所以称为支持向量机，是因为确定一个分隔超平面的不是所有的点，而是与超平面最近的若干点，这些点称为“ 支持向量 ”(空间中的点都是向量)，这样就有了支持向量机的名称。支持向量机主要是为了数量型自变量设计的。得出支持向量机回归的误判率为：0.6418383。

4.随机森林

随机森林对于大的数据库很有效率,它不惧怕很大的维数,即使是数千变量,它也不必删除变量，只要计算机能够承担，变量多多益善。得出随机森林回归的误判率为：0.01496312。它的变量重要性图（图 16）如下所示：

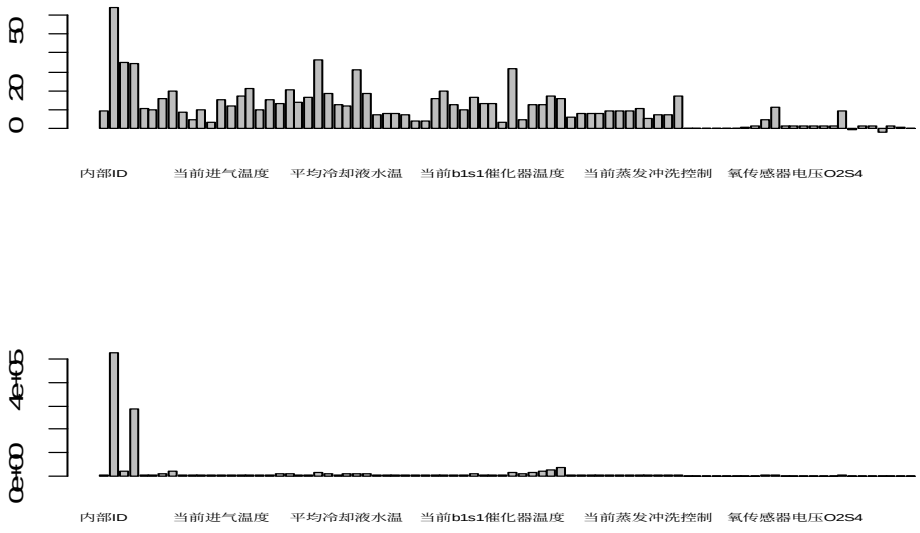


图 16 变量重要性图

从变量重要性图中可以看出重要的变量,比如最小油耗,最大油耗,剩余油量,最大进气温度,最大空气流量,最小电瓶电压,最小冷却液水温,最小 b1s1 催化器温度。

5.交叉验证比较各个模型

对于一个数据，可能有很多模型来拟合，如何衡量和比较模型预测精度？最客观的方法是交叉验证。交叉验证不需要对任何背景分布等未知的因素做任何的

假设。仅仅是用训练集训练出来的模型来预测没有用来建模的数据（测试集）。这样得出的误差是任何没有学过经典统计的人都能理解的。交叉验证可以比较任何模型，无论是经典的还是现代的。

下面是几种模型的 10 折交叉验证中对训练集预测的标准化均方误差（NMSE），结果列在下表中（表 2）：

表 210 折交叉验证训练集的 NMSE

模型	NMSE
决策树	0.2256199
boosting	0.1493707
随机森林	0. 0159227
支持向量机	0.7518383

由上表可看出：随机森林的 NMSE 最小，boosting 次之。支持向量机误差最大。

（五）多元线性模型

通过随机森林的方法，我们从 85 个变量中发现了一些相对重要变量，通过这些变量，我们以平均油耗为因变量，剩余的变量为自变量，做出来的模型并不显著，后面我们怀疑是不是存在多重共线性并做了相关检验后发现确实如此，通过消除多重共线性，优化模型。由于，最大油耗和最小油耗与平均油耗属于同一类指标，所以，我们不再考虑这两个指标。最后我们选用了如下自变量：剩余油量(X_1),最大进气温度(X_2)，最大空气流量(X_3)，最小电瓶电压(X_4)，最小冷却液水温(X_5)以及最小 b1s1 催化器温度(X_6)这 6 个变量。得到了线性回归模型,(为了方便起见，我们用X_1，…，X_6分别代表了这些变量。

$$Y = 5.9288989 + 2.3813204 * X_1 - 0.0837897 * X_2 + 0.0197739 * X_3 + -0.5318856 * X_4 + -0.0449566 * X_5 + 0.0001686 * X_6$$

由 R 运行结果可是X_1，…，X_6的系数都非常显著

由 Adjusted R-squared: 0.6537F-statistic: 3936 on 6 and 12508 DF, p-value: < 2.2e-16

修正后的 R^2 虽没有预想的那么高，但可以接受，回归方程也是比较显著的。

下面附上拟合值与残差图（图 17）,可见大部分的拟合值的残差都分布在 0 的附近。

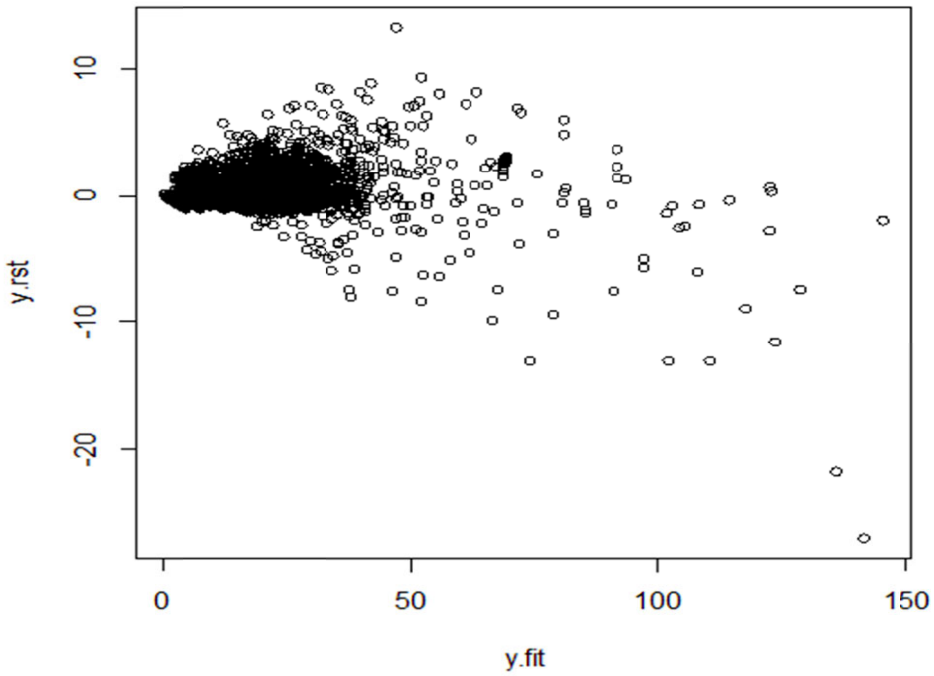


图 17 拟合值与残差图

五、主要结论和不足之处

本文通过数据挖掘和机器算法，我们从 OBD 数据集中得出这样一些结论，总体来说，我们将数据分为 8 类，代表这八种车型的发动机性能，8 号车型（长安 SC7150G 轿车）与 5 号车型（长安 SC7134C 轿车）分为一类，3 号车型（奇瑞 SQR7151A217）与 7 号车型（长安 SC7139A4B）分为一类，9 号车型(长安 SC7150G 轿车)与其他 8 种车型分为两类。聚类的效果较为理想。

通过主成分分析，随机森林等等方法的结果，我们可以知道油耗与其他变量之间存在相关关系，我们将 9 类车分成了 8 大类，通过变量提取，我们从 85 个变量中提取了 8 个变量，得到油耗的线性拟合模型，发现油耗与最大进气温度，最大空气流量，最小冷却液温度等变量有关。

本文还存在很多不足之处，数据集的数据量大，同时专业性强，我们很难判断哪些指标能删减，哪些指标应保留，我们预先删去了一些指标，可能我们数据处理的不是很好，做出来的效果不是很满意，但也能反映一些问题，希望以后能得到改进。需要后续继续研究，我们的目的是为了找出 OBD 诊断数据中，哪些参数能很好地概括并反映汽车故障信息以及油耗对尾气排放的影响，从而能让车主及时了解爱车状况，及时检修，同时也能减少汽车尾气排放。但由于时间紧迫、数据专业性太强，以及个人能力的限制，有些问题需进一步的研究。比如说，我们未能完全剔除掉不相关的指标，导致结果不是很准确。总而言之，模型建立的结果不是很符合我们的预期。

六、政策建议

我们由模型给出建议：为了响应环保节能型汽车的号召，建议汽车制造商在汽车的制造过程中应该注意上述指标的控制，车主也应该由车载诊断系统 OBD 的相关指标和我们建立的模型及时发现尾气排放的异常。我们建议在政策方面可以向汽车环保方面倾斜。在车辆的环保方面，则需要有关部门如交通部，环保局，车检部门等密切注意上述 6 个指标。实际操作方面，可以在车辆年检的时候，检测这些指标是否达标，如果不能达标，则需要勒令车主进行检修，直到排放指标正常。同时，交通部也有权对车辆进行随机抽查，并对 6 项指标进行检测，若不合格，则可以给出相应的处罚。此外，汽车制造商也需要根据这些指标对相应的零件进行检测，若发现零件不合格，工商质检部分可以勒令汽车制造商召回问题车辆。

七、附录

附录 A.数据分析的 R 软件代码；
附录 B.原始数据
由于数据量大，只列出数据集的前 10 行和后 10 行，综合参数只列出了部分变量。

表 3　即时参数数据

内部ID	平均速度	最小速度	最大熟读	当前速度	当前里程(	总里程	当前引擎转	当前引擎温	前一次时间	前一次时间	前一次时间	引擎负荷
352016850081381	21	8	68	60	1	8	2139	80	2787	2139	198	33
352016850081381	24	8	68	58	2	9	2081	82	2787	2081	198	70
352016850081381	24	5	74	30	10	11	1834	100	2678	1834	442	35
352016850081381	24	5	74	30	10	11	2058	101	2678	2058	442	29
352016850081381	24	5	74	32	10	11	2102	101	2678	2102	442	29
352016850081381	24	5	74	33	10	11	2153	101	2678	2153	442	28
352016850081381	24	5	74	20	10	11	1603	101	2678	1603	442	30
352016850081381	24	5	74	17	10	11	1366	101	2678	1366	442	33
352016850081381	24	5	74	17	10	11	1062	102	2678	1062	442	40
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
352016859398034	11	5	37	30	10	10	1269	90	1990	1269	708	7
352016859398034	11	5	37	32	10	10	1660	90	1990	1660	708	8
352016859398034	11	5	37	20	11	11	1416	89	1990	1416	708	7
352016859398034	11	5	37	11	11	11	982	90	1990	982	708	8
352016859398034	11	5	37	11	11	11	1043	90	1990	1043	708	5
352016859398034	11	1	37	1	11	11	761	90	1990	761	708	5
352016859398034	11	1	37	1	11	11	741	91	1990	741	708	5
352016859398034	11	4	37	4	11	11	769	92	1990	769	708	7
352016859398034	11	4	37	16	11	11	1488	92	1990	1488	708	14
352016859398034	11	4	37	29	11	11	1762	91	1990	1762	708	9

表 4　综合参数数据

内部ID	平均油耗	最小油耗	最大油耗	剩余油量	当前点火正	平均点火正	……	氧传感器电	氧传感器电	氧传感器电	氧传感器驾	氧传感器驾	氧传感器驾	氧传感器驾	氧传感器驾
1	11.03	7.99	60.55	7	35	6.5	……	0.82	0	0	0	0	0	0	0
1	9.28	8.43	57.12	7	7	-4.5	……	0.74	0	0	0	0	0	0	0
1	9.12	8.27	57.12	7	-9	-5	……	0.71	0	0	0	0	0	0	0
1	22.33	16.51	65.48	10	2.5	-1.5	……	0.58	0	0	0	0	0	0	0
1	12.04	8.56	60.55	7	43	28.5	……	0.68	0	0	0	0	0	0	0
1	9.1	0.25	13.41	7	-6	0	……	0.68	0	0	0	0	0	0	0
1	8.49	0.25	13.41	7	7	9	……	0.74	0	0	0	0	0	0	0
1	14.98	11.9	60.55	8	17	20.5	……	0.25	0	0	0	0	0	0	0
1	7.35	26.61	37.01	14	5.5	-6	……	0.43	0	0	0	0	0	0	0
……	……	……	……	……	……	……	……	……	……	……	……	……	……	……	……
9	17.74	17.3	1.99	6	-5	-4	……	0.72	0	0	0	0	0	0	0
3	21.99	14.39	-0.01	9	-2.5	4	……	0.11	0	0	0	0	0	0	0
3	20.84	14.39	-0.01	10	10.5	8.5	……	0.04	0	0	0	0	0	0	0
3	15.92	14.39	-0.01	10	10.5	9	……	0.03	0	0	0	0	0	0	0
3	12.96	10.33	76.05	8	5	8.5	……	0.8	0	0	0	0	0	0	0
3	12.67	10.31	76.05	8	5	8.5	……	0.12	0	0	0	0	0	0	0
3	12.43	10.26	76.05	8	9.5	10.5	……	0.63	0	0	0	0	0	0	0
3	10.24	10.15	76.05	7	9.5	8	……	0.1	0	0	0	0	0	0	0
3	10.23	10.15	76.05	8	-0.5	5.5	……	0.15	0	0	0	0	0	0	0
3	10.21	10.06	76.05	7	1	3	……	0.07	0	0	0	0	0	0	0

八、参考文献

[1] 吴喜之 ．从数据到结论[M]
[2] 吴喜之 ．复杂数据统计方法--基于 R 的应用[M]
[3] 薛薇 ．　基于 R 的统计分析与数据挖掘[M]
[4] 费宇 郭民之 ．　多元统计分析--基于 R[M]
[5] 李诗羽 王正林 ．R 语言分析:R 语言实战[M]
[6] 方匡南　朱建平 姜叶飞 ．R 数据分析方法与案例详解[M]
[7] 王敏 杨文峰 ．OBD 车载诊断系统简介[J]．汽车运用 ,2007 年 ·第 6 期 ·总第 176 期