产品预测任务

任务描述

本次比赛主要是一个对进出口交易记录数据进行产品判别的任务。本次任务有 19046 条数据记录，其中的 18279 条记录是有类别属性的，可作为分析时的训练样本，而任务目标是对 767 条测试数据（即验证样本）进行判别。

1) 数据描述

已分类的训练样本提供在比赛题目下 Excel 附件中的 cck_train 表中，训练样本的详情如下，其中，表格中的每条记录包含 7 个字段。

Number	Enterprise	Destination	Quality	Price	Origin	Custom	Product
55	1301930930	D043	17280.00	3.5000	OR028	C18	P292
56	1301930930	D019	74880.00	3.3000	OR028	C18	P185
57	1301930930	D120	42768.00	3.5500	OR028	C18	P292
58	1301930930	D111	17280.00	3.1500	OR028	C18	P292
59	1301930930	D019	187200.00	3.3000	OR028	C18	P185
60	1301930930	D110	105600.00	4.6000	OR028	C18	P185
61	1301930930	D019	149760.00	3.3000	OR028	C18	P185
62	1301930930	D079	101280.00	4.2500	OR028	C18	P185
63	1301930930	D079	101280.00	4.2500	OR028	C18	P185
64	1301930930	D019	149760.00	3.3000	OR028	C18	P185

未分类的验证样本提供在比赛题目下 Excel 附件中的 cck_test 表中。验证样本的信息如下，表格中的每条记录包含 5 个已知属性字段，其中表中属性内容与 cck_表略有不同，具体属性字段的含义请参考下节描述。

Number	Enterprise	Destination	Price	Origin	Custom	Product1	Product2	Product3
112	3301937343	D134	3.1500	OR119	C16			
113	3301969177	D094	2.5323	OR119	C16			
114	3302910017	D116	2.4174	OR122	C16			
115	3302910017	D112	3.1500	OR122	C16			
116	3302966150	D077	6.6000	OR122	C16			
117	3302966707	D019	3.7800	OR122	C16			
118	3302966707	D019	3.7700	OR122	C16			
119	3305960290	D092	2.5091	OR120	C16			
120	3308930205	D094	2.7039	OR123	C14			
121	3401960866	D119	4.2000	OR005	C16			
122	3701961548	D012	4.0550	OR082	C15			
123	3702965185	D133	3.7200	OR084	C16			

2) 属性描述

本次任务提供的样本数据包含 7 个基础属性字段，其中有 2 个连续型数值类属性字段为：Quality and Price，5 个离散型数值类属性字段为：Enterprise（560）、Destination（144）、Origin（131）、Custom（20）、Product（364）。各字段具体含义如下：

Quality:表示每条交易记录中交易产品的数量，可忽略单位。

Price:表示每条交易记录中交易产品的平均价格，单位为元。

Enterprise（560）:表示每条交易记录中交易产品的供应商编码。

Destination（144）:表示每条交易记录中交易产品的买方国家编码。

Origin（131）：表示每条交易记录中交易产品的原产地编码。

Custom（20）：表示每条交易记录中交易产品通关海关编码。

Product（364）：表示每条交易记录中交易产品的名称类别。

在验证样本中的字段 Product1 ,Product2 ,Product3 为参赛者进行分类预测后概率由大到小排名前 3 名的产品类别，字段编码同 Product 字段。

3) 样本描述

不论是在训练样本还是验证样本中，我们可以看到，一条交易记录数据包括 Enterprise（560）、Destination（144）、Origin（131）、Custom（20）、Product（364）5 个基本属性字段，括号内为每个属性下包含的所有特征值个数，而这些属性将是我们学习训练样本得到分类模型的关键，根据一条交易记录的每个属性的特征值的出现情况，利用模型对验证样本的交易产品类别进行分类预测。

结果评价

在整个验证样本预测结果中，参赛者在第 i 条记录的产品类别预测值与实际类别完全一致时可得 10 分，即预测结果字段 Product1 为实际产品类别。产品类别预测值与实际类别不一致时，其中如果预测结果 Product2 为实际产品类别的，参赛者在该条验证样本可得 2 分；如果预测结果 Product3 为实际产品类别的，该条验证样本可得 1 分，对整个 767 条验证样本预测结果加总得到一个总分 S：
取 F=S/P*100%
（其中 P 为所有验证样本类别预测结果均与实际结果相一致的总成绩，即 P=7670）
为每位参与者的模型评价得分，各位参与者模型得分由高到低依次排列。

提交说明

每位参赛者提交的结果内容如下：1、验证样本的产品类别预测结果 2、分类模型介绍（包括参数设置、假设前提等）。以上结果资料请每位参赛者者在截止日期之前以 Excel 和 PDF 形式同事发送至 ola.cheng@kcomber.com，邮件主题应该包括“队伍名称+参赛主题”。Excel 表格内容请参照训练样本格式，PDF 文件请简介明了地介绍您所使用的模型，页数不超过 4 页。

对评测任务有任何疑问，请发送邮件至 ola.cheng@kcomber.com

比赛奖金

一等奖: 1,000 RMB
二等奖: 500 RMB
三等奖: 200 RMB

重要事件节点

2016 年 5 月 20 日，发布训练数据

2016 年 8 月 8 日，发布测试数据

2016 年 8 月 15 日，评测完成，停止接受结果