机票关键词价值的预测

引言

航空公司购买搜索引擎提供的搜索关键字是一种新型商业广告的方式，可以直接提高公司的点击量从而促进机票的销售以达到盈利目的。本文通过既有的航线关键词的盈亏数据建立盈亏价值与航线信息、关键字的模型。

首先我们通过分词算法对原始词条数据提取若干的关键字特征，并对关键词的价值分为4个水平。我们基于VDM距离和卡方检验p值的思想定义了不同类之间的距离，再通过加权得到一个综合的距离。通过这一距离定义，利用改进的knn算法，采取先分类、再利用类内均值做预测得到航线价值的拟合值。

通过留一交叉验证对KNN算法的临域大小k和距离权重alpha做出估计，并评价比较模型的优劣。

为了实现模型的可泛化性，我们引入了引力模型提供对航线的另一种刻画。

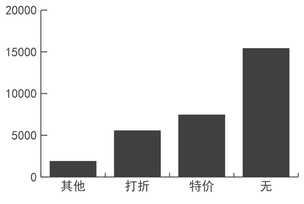
数据处理 → 距离定义 → KNN分类及预测 → 结果 → 模型的泛化

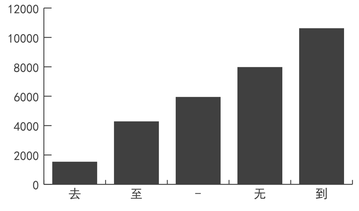
数据处理

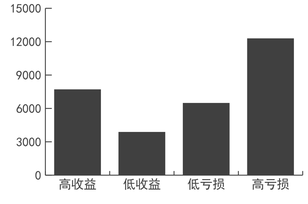
首先，对原始数据进行分词处理，整理成可用的表格结构。使用R中的Rwordseg程序包对关键词分词，分离出V1、V2、V3、V4四个特征变量。其中，V1表示出发地，V2表示目的地，V3表示动词（“到”＝1；“去”＝2；“至”＝3；“-”＝4；无＝5），V4表示形容词（“打折”＝1；“特价”＝2；“其他”＝3；无＝4）。其他表述的频数不超过总样本的5%，由于数量较少不做考虑。

关键词	利润
乌鲁木齐-阿克苏-机票	14.12
乌鲁木齐阿克苏飞机票价	9.06
乌鲁木齐到阿克苏-机票	-1.18
乌鲁木齐到阿克苏打折机票	-0.48
乌鲁木齐到阿克苏机票	31.94
乌鲁木齐-阿勒泰-机票	-1.14
乌鲁木齐-阿勒泰-特价机票	-0.49
乌鲁木齐阿勒泰订机票	9.58
乌鲁木齐阿勒泰飞机票	-0.49

列1	V1	V2	V3	V4	V5
1	乌鲁木齐	阿克苏	4	4	14.12
2	乌鲁木齐	阿克苏	4	5	9.06
3	乌鲁木齐	阿克苏	4	1	-1.18
4	乌鲁木齐	阿克苏	1	1	-0.48
5	乌鲁木齐	阿克苏	4	1	31.94
6	乌鲁木齐	阿勒泰	4	4	-1.14
7	乌鲁木齐	阿勒泰	2	4	-0.49
8	乌鲁木齐	阿勒泰	4	5	9.58
9	乌鲁木齐	阿勒泰	4	5	-0.49
10	乌鲁木齐	阿勒泰	4	1	-3.36







对于关键词价值的分类我们定义为四个小类：高收益、低收益、低亏损、高亏损。定义：y>10的关键词价值为高收益；0<y<10的关键词价值为低收益；-0.8<y<0的关键词价值为低亏损；y<-0.8的关键词价值为高亏损。

数据来源

collected by Prof. Hansheng Wang in Guanghua Business School at PKU
http://math.stanford.edu/~yuany/course/data/SE.csv

距离定义

接着，定义每个样本之间的距离，传统的距离表达显然在分类变量之间是难以完美使用的。因此，我们使用特殊的距离定义来进行衡量，使得分类之间的距离更加科学可控，在之后的KNN分类和预测以及之后泛化模型中这个距离将会被反复用到。

Distance Measure

由于我们的自变量均为多分类的定性变量，因此常规的明科夫斯基距离并不是最佳选择。对于类间距离的刻画，Stanfill和Waltz（1986）提出的VDM[①](The Value Difference Metric）是一种可行的方案。

$$vdm_a(x,y)=\sum_{c=1}^{C}\left|\frac{N_{a,x,c}}{N_{a,x}}-\frac{N_{a,y,c}}{N_{a,y}}\right|^2=\sum_{c=1}^{C}|P_{a,x,c}-P_{a,y,c}|^2$$

N_(a,x,c)表变量a取值x,输出结果为c，N_(a,x)表变量a取值x的个数

但是这一距离的定义在有些类的样本量较小时，会由于随机性的影响而缺乏稳健性，如a为航线变量时，Na,x,c的值对于很多x是非零即一的。一个自然的考虑是把VDM定义中的和式转化为加权和。而卡方检验的统计量正是一种很符合直观的加权和。如下式是检验变量a取x或y时，输出c是否服从相同分布的卡方检验的统计量。

$$\chi^2=\sum_{c=1}^{C}\frac{(P_{a,x,c}-P_{a,y,c})^2}{P_{a,x,c}}$$

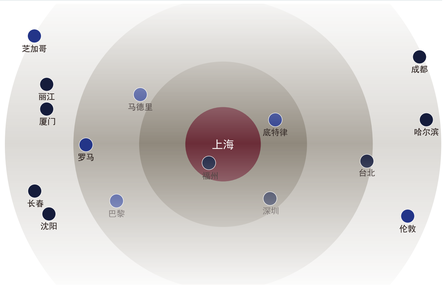
设该卡方检验的P值为Pxy，本文定义a（a=1,2,3）变量的类x和类y间的距离r_a为ln(1/Pxy)。两条关键词记录A和B之间的距离定义为r_(A,B)，关于三个变量的AB对应类的类间距离的加权平均，即

$$r_{A,B}=alpha\cdot r_{1,A,B}+r_{2,A,B}+r_{3,A,B}$$

加权的必要性来自于两方面的考量，一是不同自变量对因变量的影响程度不同，二是由于自变量不同的数据结构，它们的类间距离虽然基于同样的定义，但尺度差别较大。由于刻画文本信息的两个变量具有类似的性质，本文为了简化，假设他们权值相同，从而只对航线变量的距离赋予权值alpha。

右图所显示的是以北京为出发地的各个航线之间的距离，同心圆圆心表示北京到上海的航线，其他小圆表示北京到各个城市的航线，小圆圆心到同心圆圆心的距离表示航线之间的距离。

可以看到，在我们的距离定义下，到达丽江和厦门的航线距离接近，到达长春和沈阳的航线距离接近。



KNN分类及预测

我们主要基于k近邻模型的框架进行预测。在一般的用于预测的knn算法中，往往直接使用邻域内输出的平均值作为预测。但考虑到平均值这一度量并不稳健，且本文处理的数据方差很大，我们使用了先分类再取类内均值作为预测的策略（类似于取众数作为度量）。即先按用于分类的knn算法的思路，对某条关键词记录的价值水平做出判别，如判为“+2”，然后预测时取邻域内水平为“+2”的点的价值做平均。

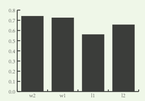
在模型中有两个参数需要设定，邻域大小k和距离权重alpha。其中k∈{30,40,50,80}，alpha∈{80,100,125}，然后通过部分数据集上的留一交叉验证进行选择。评价标准是两方面的综合考量，一方面我们希望有更好的拟合精度，即尽可能小的MSE，另一方面我们希望避免严重误判，即降低将盈利（亏损）关键词预测为亏损（盈利）的概率。根据交叉验证的结果我们选择邻域大小k为80，权重alpha为125。

模型表现

由于数据量较大，我们通过在一个较小的样本子集上进行留一交叉验证来评估模型。

如前所述，我们综合两方面的考量评价模型的表现。

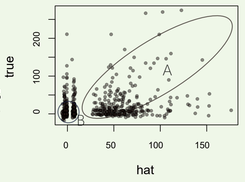
一方面是KNN模型对关键词是否盈利的预测正确率，模型表现如右图所示，效果很不错，平均正确率在70%左右。



另一方面，作为一个预测问题，我们考虑模型拟合的均方误差。交叉验证均方误差为1146.449。作为对比，我们在右侧列出了之前研究者[②]的结果，可以看出本文的KNN模型实现了更优的预测效果。

MODELS	MSE
Lasso	2001.352
Ridge	1987.391
Neural networks	2392.18

右图是交叉验证的拟合值和真实值的散点图。A区域表示对于明显盈利的关键词，模型做出了相对准确的拟合。由于我们的预测是基于分类的，因此在原本便方差较小的区间，拟合值会相对集中，即图中的B区域。在B区域内，点聚集的中心也是沿图中对角线方向分布的，这也验证了模型的有效性。·



泛化与引力模型

在现有的模型中，我们用航线间关键词价值水平分布的差异定义航线的距离，但是这一定义面对外样本（全新的航线）时缺乏泛化能力，因此我们需要建立一个模型，使得我们能够通过外部信息拟合这里定义的航线间距离。

引力模型和它的变体在多种学科领域中应用非常广泛，我们通过外部信息对航线的刻画也基于这一思想。

我们定义航线i的引力值为

$$G_i=g_{k_i}\cdot GDP_{i_1}\cdot GDP_{i_2}\cdot d_i^{p_{k_i}}$$

i 为航线编号

$$k_i=\begin{cases}1, & i\text{ 为省内航线}\\ 2, & i\text{ 为跨省航线}\\ 3, & i\text{ 的目的地或出发地为境外城市}\\ 4, & i\text{ 跨境，但无明确地目的地城市}\end{cases}$$

$GDP_{i_1}GDP_{i_2}$ 为出发地和目的地的人均GDP（人民币）

为航线的直线里程(km)
为待估参数

我们希望用 $\frac{G_i\wedge G_j}{G_i\vee G_j}$ 来拟合　(前述卡方检验的p值)

$$\because r_{i,j}=-\log(P_{i,j})$$

$$\therefore r_{i,j}\sim\log(G_i\vee G_j)-\log(G_i\vee G_j)$$

$$=|\log G_i-\log G_j|$$

$$=|C_{k_i}-C_{k_j}+log\frac{GDP_{i_1}\cdot GDP_{i_2}}{GDP_{j_1}\cdot GDP_{j_2}}+p_{k_i}\cdot logd_i-p_{k_j}\cdot logd_j|\triangleq f(i,j)$$

$\{g_k\}\{p_k\}$ 为 $\text{s.t.}\sum_{i,j}(r_{i,j}^2-f^2(i,j))^2$ 最小化的参数

不过，引力模型拟合的效果并不十分理想。这可能是因为我们考虑的航线信息不够或者引力模型的函数形式有误，因此现阶段模型的泛化能力有限，不足以很好的处理对全新航线的预测，有待我们进一步的研究。

小组成员

张栩川（1501210027）：数据预处理、KNN模型
彭俊菁（1501210045）：数据预处理、引力模型
杨新宇（1501210053）：KNN模型、距离刻画
张立（1501210057）：海报制作、数据预处理
张琪（1501210058）：距离刻画、引力模型

参考文献

① Stanfill, C., and D. Waltz, (1986). Toward memory-based reasoning. Communications of the ACM, Vol. 29, December 1986, pp. 1213-1228.
② http://math.stanford.edu/~yuany/report/Poster06.Keyword.pdf