数据挖掘大作业结题报告

李苏畅 李杨 罗伟 魏思杰

## **1 项目介绍**

我们组选择的项目是kaggle上的一个旅馆推荐的题目。项目目标是对已经进行过聚类处理的旅馆类别中预测用户最有可能选择的5组旅馆类别并作出推荐。数据中包含了大量的信息，包括了用户的预定时间，所使用的门户网站，用户所在的国家、城市，用户与目的地距离，用户是否使用移动设备，用户是否携带行李，用户的登入登出日期，用户包含的成人、儿童数量，用户预定的房间数，旅馆所在的大陆、国家、市场，以及用户做过几次类似的预定等数据。

项目介绍链接：

https://www.kaggle.com/c/expedia-hotel-recommendations

项目数据集链接：

<https://www.kaggle.com/c/expedia-hotel-recommendations/data>

## 2 算法简介

##### 2.1 GBDT工作原理

GBDT(Gradient Boosting Decision Tree)又叫MART(Multiple Additive Regression Tree)，是一种迭代的决策树算法，该算法由多棵决策树组成，所有树的结论累加起来做最终答案。它在被提出之初就和SVM一起被认为是泛化能力(generalization)较强的算法。近些年更因为被用于搜索排序的机器学习模型而引起大家关注。

GBDT几乎可用于所有回归问题（线性/非线性），相对logistic regression仅能用于线性回归，GBDT的适用面非常广。GBDT亦可用于二分类问题（设定阈值，大于阈值为正例，反之为负例）。

GBDT主要由三个概念组成：Regression Decision Tree（即DT），Gradient Boosting（即GB），Shrinkage（算法的一个重要演进分枝，目前大部分源码都按该版本实现）。

###### 2.1.1 回归树(Regression Decision Tree)

决策树分为两大类，回归树和分类树。前者用于预测实数值，如明天的温度、用户的年龄、网页的相关程度；后者用于分类标签值，如晴天/阴天/雾/雨、用户性别、网页是否是垃圾页面等。前者的结果加减是有意义的，如10岁+5岁-3岁=12岁，后者则无意义，如男+男+女没有意义。GBDT的核心在于累加所有树的结果作为最终结果，就像前面对年龄的累加，而分类树的结果显然是没办法累加的，所以GBDT中的树都是回归树，不是分类树。

以对人的性别判别/年龄预测为例，每个instance都是一个我们已知性别/年龄的人，而feature则包括这个人上网的时长、上网的时段、网购所花的金额等。

C4.5分类树在每次分枝时，穷举每一个feature的每一个阈值，找到使得按照feature阈值和feature>阈值分成的、两个分枝的熵最大的feature和阈值（熵最大的概念可理解成尽可能每个分枝的男女比例都远离1:1），按照该标准分枝得到两个新节点，用同样方法继续分枝直到所有人都被分入性别唯一的叶子节点，或达到预设的终止条件。若最终叶子节点中的性别不唯一，则以多数人的性别作为该叶子节点的性别。

回归树总体流程也是类似，不过在每个节点（不一定是叶子节点）都会得一个预测值，以年龄为例，该预测值等于属于这个节点的所有人年龄的平均值。分枝时穷举每一个feature的每个阈值找最好的分割点，但衡量最好的标准不再是最大熵，而是最小化均方差，即，或者说是。这很好理解，被预测出错的人数越多，错的越离谱，均方差就越大，因此可以通过最小化均方差来找到最靠谱的分枝依据。分枝直到每个叶子节点上人的年龄都唯一或者达到预设的终止条件（如叶子个数上限），若最终叶子节点上人的年龄不唯一，则以该节点上所有人的平均年龄作为该叶子节点的预测年龄。

###### 2.1.2 梯度迭代(Gradient Boosting)

Boosting，迭代，即通过迭代多棵树来共同决策。GBDT是把所有树的结论累加起来做最终结论的，所以可以想到每棵树的结论并不是年龄本身，而是年龄的一个累加量。GBDT的核心就在于，每一棵树学习的是之前所有树结论和的残差，这个残差就是一个加入预测值后可以得到真实值的累加量。举例说明：假如A的真实年龄是18岁，但第一棵树的预测年龄是12岁，即残差为6岁。那么在第二棵树里我们把A的年龄设为6岁去学习。如果第二棵树真的能把A分到6岁的叶子节点，那累加两棵树的结论就是A的真实年龄；如果第二棵树的结论是5岁，则A仍然存在1岁的残差，第三棵树里A的年龄就变成1岁，然后继续学习。这就是梯度迭代在GBDT中的意义。

###### 2.1.3 缩减(Shrinkage)

Shrinkage的思想认为，每次走一小步逐渐逼近结果的效果，要比每次迈一大步很快逼近结果的方式更容易避免过拟合。即它不完全信任每一个棵残差树，它认为每棵树只学到了真理的一小部分，累加的时候只累加一小部分，然后通过多学几棵树弥补不足。

经验证明，和Adaboost一样，Shrinkage也能减少过拟合的发生。

##### 2.2 PCA工作原理

对同一个体进行多项观察时，必定涉及多个随机变量*X*1，*X*2，…，*Xp*，它们都是的相关性，一时难以综合。这时就需要借助主成分分析(principal component analysis)来概括诸多信息的主要方面。我们希望有一个或几个较好的综合指标来概括信息，而且希望综合指标互相独立地各代表某一方面的性质。

任何一个度量指标的好坏除了可靠、真实之外，还必须能充分反映个体间的变异。如果有一项指标，不同个体的取值都大同小异，那么该指标不能用来区分不同的个体。由这一点来看，一项指标在个体间的变异越大越好。因此我们把“变异大”作为“好”的标准来寻求综合指标。

###### 2.2.1主成分的一般定义

设有随机变量*X*1，*X*2，…，*Xp*，其样本均数记为，，…，，样本标准差记为*S*1，*S*2，…，*Sp*。首先作标准化变换

我们有如下的定义：

(1)若*C*1=*a*11*x*1+*a*12*x*2+…+*a*1*pxp*，，且使*Var*(*C*1)最大，则称*C*1为第一主成分；

(2)若*C*2=*a*21*x*1+*a*22*x*2+…+*a*2*pxp*，，(*a*21，*a*22，…，*a*2*p*)垂直于(*a*11，*a*12，…，*a*1*p*)，且使*Var*(*C*2)最大，则称*C*2为第二主成分；

(3)类似地，可有第三、四、五…主成分，至多有*p*个。

###### 2.2.2 主成分的性质

主成分*C*1，*C*2，…，*Cp*具有如下几个性质：

(1)主成分间互不相关，即对任意*i*和*j*，*Ci*和*Cj*的相关系数

*Corr*(*Ci*，*Cj*)=0 *ij*

(2)组合系数(*ai*1，*ai*2，…，*aip*)构成的向量为单位向量，

(3)各主成分的方差是依次递减的，即

*Var*(*C*1)≥*Var*(*C*2)≥…≥*Var*(*Cp*)

(4)总方差不增不减， 即

*Var*(*C*1)+*Var*(*C*2)+…+*Var*(*Cp*)=*Var*(*x*1)+*Var*(*x*2)+…+*Var*(*xp*)

这一性质说明，主成分是原变量的线性组合，是对原变量信息的一种改组，主成分不增加总信息量，也不减少总信息量。

(5)主成分和原变量的相关系数*Corr*(*Ci*，*xj*)=*aij*=*aij*

(6)令*X*1，*X*2，…，*Xp*的相关矩阵为*R*，(*ai*1，*ai*2，…，*aip*)则是相关矩阵*R*的第*i*个特征向量(eigenvector)。而且，特征值*i*就是第*i*主成分的方差，即

*Var*(*Ci*)=*i*

其中*i*为相关矩阵*R*的第*i*个特征值(eigenvalue) **1≥**2≥…≥*p*≥0

##### 2.3.主成分的数目的选取

前已指出，设有*p*个随机变量，便有*p*个主成分。由于总方差不增不减，*C*1，*C*2等前几个综合变量的方差较大，而*Cp*，*Cp*-1等后几个综合变量的方差较小，严格说来，只有前几个综合变量才称得上主（要）成份，后几个综合变量实为“次”（要）成份。实践中总是保留前几个，忽略后几个。

保留多少个主成分取决于保留部分的累积方差在方差总和中所占百分比（即累计贡献率），它标志着前几个主成分概括信息之多寡。实践中，粗略规定一个百分比便可决定保留几个主成分；如果多留一个主成分，累积方差增加无几，便不再多留。

##### 2.4.主成分回归

主成分分析本身往往并不是目的，而是达到目的的一种手段。因此，它多用在大型研究项目的某个中间环节。例如，把它用在多重回归中，便产生了主成分回归。另外，它还可以用于聚类、判别分析等。本节主要介绍主成分回归。

在多重回归曾指出，当自变量间高度相关时，某些回归参数的估计值极不稳定，甚至出现有悖常理、难以解释的情形。这时，可先采用主成分分析产生若干主成分，它们必定会将相关性较强的变量综合在同一个主成分中，而不同的主成分又是互相独立的。只要多保留几个主成分，原变量的信息不致过多损失。然后，以这些主成分为自变量进行多重回归就不会再出现共线性的困扰。如果原有*p*个自变量*X*1，*X*2，…，*Xp*，那么，采用全部*p*个主成分所作回归完全等价于直接对原变量的回归；采用一部分主成分所作回归虽不完全等价于对原变量的回归，但往往能摆脱某些虚假信息，而出现较合理的结果。

以上思路也适用于判别分析，当自变量高度相关时，直接作判别分析同样有多重共线性问题，可先计算自变量的主成分，然后通过主成分估计判别函数。

## 3.计算结果并输出

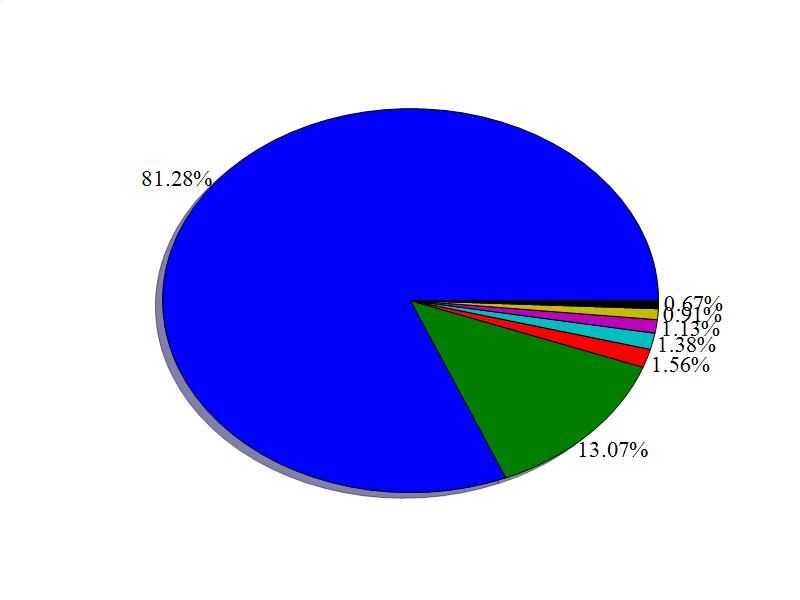
由于数据集是随机抽取并保护了用户信息的数据集，所以其中存在大量的数据缺失，由图1-1可得，数据项之间的关联程度很低，所以我们为了更准确地进行数据分析，首先对数据进行了清洗，将数据集分成了多个子数据集。个别缺失数据较多的数据集就被放弃了。训练数据集中共涉及120万用户，旅馆类别100类。旅馆类别明显较少，故我们对每一类的旅馆分别进行了绘图后的数据分析。

图3-1 不同出发点的游客占旅馆总人数的比例

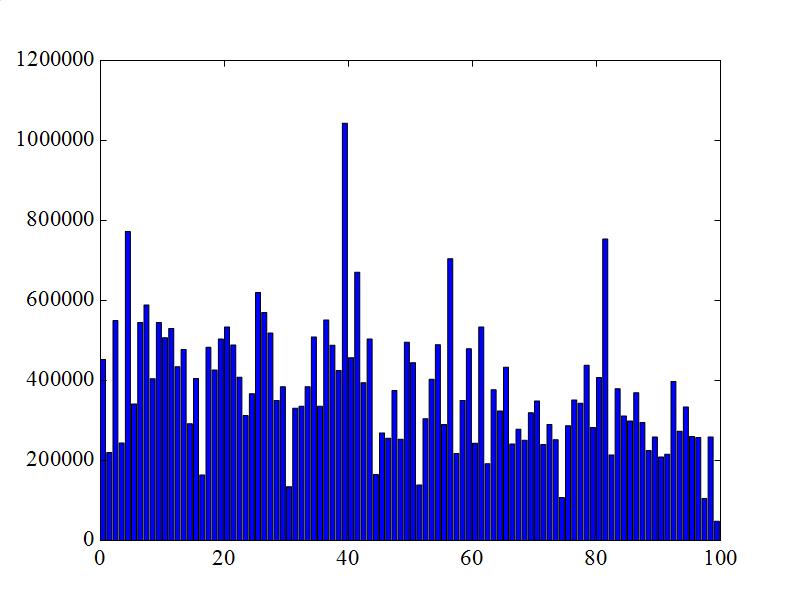
图3-2 不同旅馆的游客数比较

图3-1和图3-2是对第个类别的旅馆进行相关数据分析的截图。

通过小组成员间讨论以及本项目前中期以来的经验，我们确定了采用的算法。最终，我们使用PCA算法对数据集进行特征提取，选取了7种特征，对提取出来的特征使用GBDT回归，来拟合训练数据集的曲线，用得到的参数作为系数来计算得分。最终利用TopN算法，将得分最高的旅馆类别作为推荐的类别，直到获得5种旅馆类别。并将结果作为测试数据的结果进行输出。

## 4.实验结果和展望

本项目的得分计算公式是。其中U是用户事件的数量，P(k)是k时的预测值，n是预测的旅馆聚类的数量。根据网站给出的返回结果，我们算法的最终得分是0.50186。虽然预测的效果与最优结果相比还有一定的距离（排名大概150名），但是相比于中期时使用的纯统计学方法已经有了长足的提高。未来有机会我们会考虑尝试进一步优化特征提取的结果，并争取通过对参数的改进来提高GBDT的回归效果，希望能够获得更好的实验结果。