基于特征工程与协同过滤的餐饮推荐

**开题报告**

1. **课题背景**

随着Internet上的信息呈现指数级的增长规模，如何为用户提供准确而精简的信息、使他们不迷失在海量信息中成为了重要课题。与被动接受用户的需求，按指令与关键词提供结果的搜索引擎技术不同，推荐引擎或推荐系统则是主动从用户的历史活动数据中学习其需求，进而较为准确地预测用户的爱好。在这个新机遇的背景下，包括电影信息网站Netflix、产品点评网站Epinions和餐饮为主的点评网站Yelp在内的许多数据驱动的国外网站纷纷公开实际数据集，有奖征募数据挖掘方案，并推动推荐系统各个子领域算法的进展；例如Netflix Prize提供了协同过滤领域最全面的数据集。而本文基于的Yelp数据集则不仅包括全球各地海量的用户、餐馆和点评信息，更蕴含有用户之间直接形成的社交关系网的数据，在数据维度的精细程度上更胜一筹。Yelp数据集竞赛（Yelp Dataset Challenge）已经成功举办了六届并公布了五届的成果，通过阅读可以对餐饮数据挖掘问题的特点、难点有一个比较全面的认识。

**1.1 推荐系统基本方法介绍**

作为机器学习技术在实际生产中的一类重要应用，构成推荐系统核心的算法通常分为两大类，即基于内容的过滤（content-based filtering）与协同过滤（collaborative filtering）。

* + 1. **基于内容的过滤（简称CBF）方法**

抽取出的产品和用户的特征来分别刻画他们的档案（profile），然后通过计算用户与产品之间的匹配程度来给出适当的推荐。通常的CBF采纳用户所选择或消费的产品的统计信息来刻画用户自己的档案；而在Yelp数据集中，用户与产品（即以餐馆为主的商业机构，以下简称餐馆或者business）之间的联系要多元、复杂得多，使用各种自然语义处理（natural-language processing）手段，如话题模型（topic model）对点评（review）中的丰富信息进行挖掘是一个热点的研究角度。

* + 1. **协同过滤（简称CF）方法**

在全领域内的研究则在近年来愈见热度。它避免了依赖复杂的人工特征工程来产生刻画用户和产品的特征，只需要简单、通用地预处理数据，便可根据用户过往的消费选择及评分，获知用户与产品之间的相互作用，进而产生个性化的推荐。由此CF基本克服了CBF需要大量领域知识的问题，较容易推广，却也引入了新的问题——数据稀疏性（sparsity，或可扩展性 scalability）和冷启动（cold start）。常见的CF方法包括以下几类：

* **基于记忆的方法（memory-based approaches）**

利用统计方法从整个数据集来产生推荐和预测，一般包括三个步骤：（1）定义相似度计算方法，如皮尔逊相关度系数（Pearson Correlation Coefficients，简称PCC）和余弦相似度（Cosine Similarity）；（2）基于历史评分数据等，用K-近邻（K Nearest Neighbors，简称kNN）等算法在数据集中为用户找到相似的用户（user-based），或为产品找到相似的产品(item-based)作为近邻；（3）整合利用近邻的信息，为目标用户生成个性化推荐。

* 基于模型的方法（model-based approaches）

将某个时刻的数据集视为训练集，让基于概率矩阵分解（probabilistic matrix factorization）、隐变量（latent factor）等的模型进行学习后，再用于预测用户未来的评分，将期望评分高的产品顺理成章地推荐给用户。许多该类的方法可以被视为矩阵填补问题，在稀疏矩阵的情况下可扩展性明显好于基于记忆的方法，问题主要是隐变量模型不容易解释，以及增量学习的设计等。

**1.2 算法评估准则**

对于一般的算法，我们主要关注其运行的时间和空间复杂度等信息，来作为其优化目标；而对于推荐系统中常用的模型训练算法与机器学习分类算法而言，虽然时间空间复杂度依然重要，但需要放在以下的标准（performance metrics）之后考虑。

在基于模型的方法的研究中，测度模型或算法的质量的标准主要是均方根（root mean squared error，简称RMSE）和平均绝对误差（mean absolute error，简称MAE）。记，在我们的测试集T中，用户u对商铺b的打分被预测为，而实际上该值为，那么定义RMSE和MAE如下：

,

而对于基于特征工程和CBF等的机器学习分类方法来说，常用的调优测度标准还有准确率（Precision），召回率（Recall）和F1-Score。定义如下：

P = Precision = TP / (TP + FP), R = Recall = TP / (TP + FN)

其中TP = True Positive，FP = False Positive，FN = False Negative；事实上P和R可以基于每一个类分别定义，若设某类为A，则P = (被分为A的样本中真实为A的 / 全部被分为A的样本个数)，R = (被分为A的样本中真实为A的 / 全部实际上为A的样本个数)。

而作为一个综合性的评估准则，F1-Score 被定义为上述二者的调和平均：

F1-Score = 2PR / (P + R)

[6] 进一步提出使用残差和（Residual Sum，简称RSS）来校正Yelp五级评分中明显的训练数据偏差，正如文献综述将详细提到的那样。

**1.2 特征工程方法介绍**

许多机器学习和推荐系统相关的理论书籍中并没有一章来详细阐述特征工程的原理和方法，而事实上基于Yelp为代表的餐饮数据的独特属性，基于领域知识的特征工程方法在此类问题的推荐上经常有大有可为。所谓“数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已”。特征工程中一般包括如下几个子问题：特征提取（feature extraction）、特征构造（feature construction）和特征选择（feature selection），而它们经常被混合提及。一般来说，我们期望选择一组彼此正交（独立、关联性低）的特征，它们的线性或非线性组合与我们要预测的目标函数或属性（比如用户对产品的评分）则有较强的关联性（correlation）。

**1.2.1 特征提取和特征构建**

特征提取的对象是原始数据（raw data），它的目的是自动地构建新的特征，将原始特征转换为一组具有明显物理意义、几何特征、统计意义或核（kernel）性质的特征。矩阵或表格数据上，常用的提取方法包括主成分分析（principal component analysis，简称PCA）、独立成分分析（independent component analysis，简称ICA）和线性判别分析（linear discriminant analysis，简称LDA）；在图象识别中可能还包括线检测、边缘检测等。

而至于特征构建，指的一般是从原始数据中人工构建新的特征，与基于PCA等的自动采样不同。这需要对领域知识的深刻洞察。

**1.2.2 特征选择**

特征选择主要是从上述处理得到后的特征集合中，挑选一组最具统计意义的特征子集，从而达到降维的效果。因为原始的数据特征可能经过了PCA等处理后已经发生了维度的改变，一般需要借助自动化乃至可视化（data visualization）的工具来进行。

* 筛选器方法（filter）：我们可以使用一些评价指标单独地计算出单个特征跟类别变量之间的关系，如Pearson相关系数，Gini-index（基尼指数），IG（信息增益）等，然后再挑选相关性最高的一组特征进行训练，通常能够让我们向特征工程的目标更进一步。该类方法计算较为高效，也不容易过拟合，但倾向于忽略特征之间的相关性，也即多个特征组合起来可能几何级别提升预测效果的可能性。
* 封装器方法（wrapper）：实质上是一个分类器，封装器用选取的特征子集对样本集进行分类，分类的精度作为衡量特征子集好坏的标准，经过比较选出最好的特征子集。常用的有逐步回归（Stepwise regression）、向前选择（Forward selection）和向后选择（Backward selection）。它可以克服筛选器的特征之间关联问题，但是在样本较少时容易过拟合，而样本数据量大时计算效率不高。
* 集成方法（embedded）：让学习器使用正则化（regularization）等方法自主做特征选择，或者使用决策树思想，如工业界常用的随机森林（random forest）和迭代提升（gradient boosting）方法。

如下图所示，特征选择过程一般包括产生过程，评价函数，停止准则，验证过程，这4个部分：



**1.3 Yelp数据集的特性详解**

根据Yelp Dataset Challenge官网的说明，该数据集包括来自约552K用户对77K的商铺给出的2.2M条评价，以及591K的小贴士（tips）、大量的商铺签到、3.5M条用户之间的社交联系；除此之外，今年还新增了200K张商铺图片，鉴于这是计算机视觉（CV）等领域的研究重点，本文将并不涉及图片的处理和挖掘。接下来对经过属性预筛选的数据集的形式，和各个属性可能的利用方案做一详述。

（注意：Yelp在公布数据集的时候对用户、商铺和评论的id都做了加密处理，应该也做了不影响整体统计特征的扰动，以保护用户隐私。）

**1.3.1 Business**

1. business\_id: 加密后的商铺唯一id；
2. name: 商铺名称；
3. neighborhoods: 商铺属于的街区范围名称的列表，根据统计每个商铺可能属于0-3个街区；
4. city: 商铺所处城市；
5. state: 商铺所处的州或省的缩写；
6. stars: 1-5分离散打分制下的商铺平均得分（四舍五入到0.5分的粒度）；
7. review\_count: 商铺收到的评分总数；
8. categories: （本地化的，每个城市可能不同的）商铺类别列表；
9. open: 布尔值，如果为false则说明商铺歇业，可以忽略之；
10. weekends: 将原本的精细到一周中的每一天营业时间的`hours`粗粒度化为是否在周末营业的布尔值。

从business中剔除了详细地址信息、经度、纬度这三个地址位置相关的属性；attributes中全部的过分详细的可选商铺属性也暂时排除，如是否有车位、是否有WiFi等，因为拥有每个属性的值的商铺数量占全部商铺比例较小，在许多模型中都容易出现过度拟合的问题，在kNN中则容易找不到近邻或近邻数量太少，造成拟合不足（high bias）。

另外做初步统计的时候发现，有16个州的商铺数量不足17家，与此对应的是剩下12个州（SC，PA，ON，MLN，IL，QC，WI，BW，NV，AZ，NC，EDH）的商铺信息都在136家及以上，因此从business、review和user中剔除这16个州的全部数据，避免引入额外的误差因素。

**1.3.2 Review**

1. review\_id: 加密后的评论唯一id；
2. business\_id: 加密后的商铺唯一id；
3. user\_id: 加密后的用户唯一id；
4. stars: 该评论对商铺的评分（1-5星的整数离散值，官网的数据说明有误）；
5. date: 点评的日期；
6. votes: 点评收到的赞数总和（虽然Yelp提供了三种不同的vote选项，其中useful的总数约为cool和funny的总和，但是根据[8]的说法，经过测试发现对它们赋予不同的权，结果没有明显差异，因此只考虑总数）。

因为本文不关注NLP相关的研究，暂且不考虑收入评论的文本内容。另外录入数据时应该刨去对已关闭商户的评价。

**1.3.3 User**

1. user\_id: 加密后的用户唯一id；
2. name: 用户的名（可以用来判断用户性别）；
3. review\_count: 用户总计点评数；
4. average\_stars: 用户历史评分的平均值，浮点数；
5. votes: 简化为用户的点评收到的支持票总和；
6. friends: 用户的好友id列表，用以构建互信关系网络；
7. elite: 用户自荐成为VIP的年份列表；
8. yelping\_since: 注册日期（年和月）；
9. compliments: 包括`{'photos', 'writer', 'cool', 'list', 'profile', 'funny', 'plain', 'hot', 'note', 'more', 'cute'}`这几种，同样简化为用户收到的各种类型的赞扬数总和；
10. fans: 关注该用户的人的数量。

**一些元数据：**

|  |  |  |
| --- | --- | --- |
| 项目 | 用户数量 | 收到的平均赞扬数 |
| 总 | 552339 | 20.00 |
| VIP | 31461 | 305.64 |

上述VIP用户统计里包括所有曾经成为过Elite的用户，而他们中保有这个荣誉的平均时长为2.88年。

**1.3.4 Tip & Check-in**

小贴士（tip）是用户对某商铺给出的“轻量级的”、不带评分的点评，额外信息只有`text`和`date`。大致上每位用户平均有写过一条；经统计发现，tip的平均长度为60个字符，而tip的like功能虽然设计与review的vote功能类似，却几乎没有人使用，总计只有8155条被like过，还不到总数的1.38%，平均每条tip只被喜欢0.015次。

虽然Yelp的商铺签到（check-in）给出了精确到一周中每个小时的数据，目前认为这对于推荐系统的主要任务而言属于太细的粒度，因此暂时只计算签到总量，计入business的checkin属性。

二. 目标和任务

鉴于目前基于餐饮数据的挖掘工作多侧重NLP领域展开，而对Yelp等数据集中提供的其他丰富的属性、特征探究不足，本次研究将运用各种特征工程手段，详细研究几个潜在的关键属性对推荐结果造成的影响，并据此选择合适的机器学习算法，实现一套预测用户对商铺评分效果较为理想的CBF推荐系统。主要探究的点可能有：（1）当地用户与外地来的用户对同一类餐馆的评分差异；（2）近期热点（trending）餐厅的评分分布、变化趋势及将其推荐给新用户的效果；（3）用户撰写的餐馆评论收到的赞同投票（votes）对推荐效果的影响；等等。

同时，因CBF方法对于用户之间的社交关系网络利用不足，通常不能很好地反映出现实中熟人推荐情形的实质，我们还有必要实现[1]中基于隐变量模型和概率矩阵分解的CF算法，在不同训练数据规模、不同分类的商铺-用户组合上，与经过充分调优的、基于特征的CBF算法进行效果对比，从而明确在该类餐饮商业数据集上预测效果较好、或效率较高的推荐算法类型。

最后，可以考虑将两种算法的实现进行加权组合或或者序贯组合，提升该推荐系统的整体效果。

**三. 可行性分析**

本研究拟主要使用基于Python的各机器学习相关库进行数据清洗、存储、处理，特征工程，建模学习，以及模型组合与调优等各个环节，如NumPy、SciPy、Matplotlib、Pandas和Scikit-learn（Sklearn）等，相对开发快捷方便。

本研究拟使用的各机器学习模型，如简单贝叶斯分类器、逻辑回归、支持向量机、随机森林和GBDT都较为成熟，在Python上均有对应版本的实现，调用比较容易，可以帮助将工作专注在特征选择、模型组合与效果对比上。

**四. 研究方案和关键技术考虑**

**4.1 数据清洗与存储**

在该阶段首先使用Python 3.4.4，按照如上的数据详述，对Yelp提供的JSON格式数据统计一些元数据（metadata），简单清洗、简化后存入一个SQLite3数据库，使用Firefox上的扩展插件SQLite Manager做图形界面的辅助检查。

**4.2 数据可视化与预处理**

Pandas库在数据预处理以及后续应用于模型的诸多流程中非常有用，它构建于科学计算库NumPy之上，与推荐算法库Sklearn之间有易于使用的接口，支持以下需求：

1. 支持自动或明确的数据对齐的带有标签轴的数据结构。这可以防止由数据不对齐引起的常见错误，并可以处理不同来源的不同索引数据；
2. 整合的时间序列功能；
3. 以相同的数据结构来处理时间序列和非时间序列；
4. 支持传递元数据（坐标轴标签）的算术运算和缩减；
5. 灵活处理丢失数据；
6. 在常用的基于数据的数据库（例如基于SQL）中的合并和其它关系操作。

在数据预处理和特征工程阶段均需要一定的数据可视化技术，拟混用MatPlotLib和各基于JavaScript的可视化库或现成工具来完成，如d3.js，基于d3的RAW，百度的echarts 3和Plotly等。

**4.3 特征工程**

该步与下述4.4中的模型训练是一个循环迭代的过程，根据预测效果指标持续优化、不断提升。

通过PCC和可视化等方法先对特征对初步筛选，选择一两个简单高效的分类器（如朴素贝叶斯、逻辑回归）串联起基于Sklearn的推荐系统，快速实现基准线（baseline）模型作为后续实验的评估准则（benchmark）。首先研究在目标和任务中提出的几个新特征对结果有何增益或影响，然后逐步应用特征提取和特征选择的方法，引入性能更优良的特征来改善推荐预测的评估指标。注重该步骤工具的自动化、可迭代化，考虑使用Gradient Boosting方法。

**4.4 基于机器学习的模型训练**

该步与上述4.3中的特征过程是一个循环迭代的过程，特征过程根据模型训练得到的预测效果指标持续优化、不断提升。

因为重点在于特征的分析研究，该部分将侧重于挑选常用的、参数选择方便、训练效果较高的模型，比如先从朴素贝叶斯和逻辑回归开始，若运行效率可以接受则可以再考虑使用随机森林甚至参数选择较为复杂的SVM等。这些基础算法在Sklearn框架中都有实现版本。

**4.5 基于CF和矩阵分解的模型训练**

基于PMF等的统计学习训练模型只对数据清洗和预处理有一定要求，而不需要特征工程的介入，可以自行挖掘潜在特征空间做矩阵填补。Sklearn中实现有简单的SVD版本，而因为有[1]等一些文献提供了考虑社交网络结构的矩阵分解方法，也可以自己据此实现一个版本。在Matlab中，给予目标函数（objective function）和梯度下降（gradient descent）的表达，内置的库可以非常方便地根据限定的最大迭代次数对它们进行迭代求解，并求出效果变化的曲线；然而Matlab的运行效率一般来说在较大规模数据集上不可忍受，考虑使用基于NumPy和Matplotlib予以替代，语法类似。

**五. 预期研究结果**

本研究项目预期发掘出餐饮数据集上具有代表性的一组特征集，选择合适的机器学习方法对用户-产品评分做出预测；并实现文献中基于概率矩阵分解、考虑社交网络结构的隐因子模型，同样获得推荐结果。随后将参数调优后的两种方法进行对比、解释以至于融合。

**六. 进度计划**

* **2016-02至2016-03-30**

资料收集，文献阅读与思路整理，外文翻译，撰写文献综述与开题报告。

了解目前推荐系统领域的主要模型与方法，工业界和数据竞赛常用方法、学术界研究的热点问题；主要通过阅读Google Scholar中整理的Recommend Survey等材料，和实验室类似方向的学长、导师的指导。

完成数据的初步清洗、归档与数据库的搭建，熟悉了解项目需要使用的各类库，如Sklearn等。

* **2016-04-01至2016-04-15**

完成数据的清洗、预处理和可视化工作，对数据的特性进行深入探究，构建baseline版本，取得基于特征过程和CBF的机器学习方法的初步预测结果，并明确不足（underfit或overfit等）和优化方向。

* **2016-04-16至2016-04-30**

着重特征工程，探究餐饮数据集中一些独特的重点特征对推荐系统预测结果的影响，优化调参。

* **2016-05-01至2016-05-15**

着重基于PMF的CF方法的构建，优化调参后与前一阶段结果进行对比解释。

* **2016-05-15至2016-05-30**

融合模型，整理成果，规范化论文的撰写。