2020美赛O奖C题笔记(2004647)

2021年2月4日 9:42

1. summary：
    a. 题目背景；
    b. 模型简介：
    (名字+作用+result);
        i. RRBS(评价模型)：基于rating和review定义了客户对产品的分数。the Reputation Model(评价+时序模型)：在RRBS基础上加上时间。result：reputation和销量有正相关；
        ii. rating和review的关系：result: high rating刺激更多积极的review；low rating则不一定刺激更多负面的review。result：rating和graded words之间有正相关；
    c. strategy；
2. Introduction：
    a. 一段背景；
    b. Problem Restatement；
    c. Literature Review(这个领域过去的方法+本文的改进)：
        i. 利用更多信息：文本长度和特定单词的情感强度；
        ii. 其他的有用信息：e.g. helpfulness rating；
        iii. 提出基于review和rating组合的度量；
    d. Data Cleaning:(异常信息+处理方法+方法的合理性)
    包含错误类别的商品，都是只有个位数评论，而个位数评论对建模没有什么影响，所以删掉；
    e. Modeling Framework：一张图；

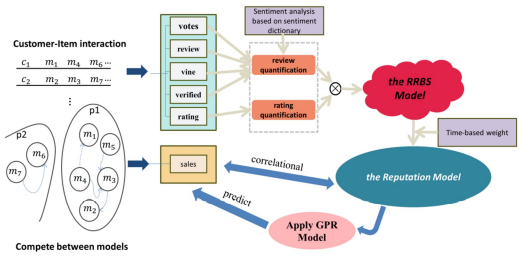

Figure 1: Modeling Framework

3. Assumptions & Nomenclature；
4. models：(用这个模型的原因+模型描述+结果(验证模型正确性))
    a. the RRBS Model(评价模型)：
    (第i种产品第k个品牌第t个月的得分)α是rating向量，β是review向量，λ是权重因子；

$$score_{i,k}^{(t)} = \boldsymbol{\alpha}^{\circ\lambda}\boldsymbol{\beta}$$

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_{i,k,1}^{(t)} & \alpha_{i,k,2}^{(t)} & \cdots & \alpha_{i,k,j}^{(t)} & \cdots & \alpha_{i,k,n}^{(t)} \end{pmatrix}$$

其中，α的值由下表确定：

Star Rating	Mapped Rating
1-star & 5-star	3
2-star & 4-star	2
3-star	1

$$\boldsymbol{\beta} = \boldsymbol{A\Phi} = \begin{pmatrix} \beta_{i,k,1}^{(t)} & \beta_{i,k,2}^{(t)} & \cdots & \beta_{i,k,j}^{(t)} & \cdots & \beta_{i,k,n}^{(t)} \end{pmatrix}$$

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_{i,k,1}^{(t)} & \phi_{i,k,2}^{(t)} & \cdots & \phi_{i,k,j}^{(t)} & \cdots & \phi_{i,k,n}^{(t)} \end{pmatrix}^T$$

$$\phi_{i,k,j}^{(t)} = \theta_j \cdot s_j^{v_j} \cdot h_j \cdot L_j$$

或者

$$\phi_{i,k,j}^{(t)} = \theta_j \cdot s_j^{v_j} \cdot e^{\frac{helpful\_votes_j}{total\_votes_j} - 0.5} \cdot \frac{1}{2}\log_{10} len_j$$

$$\theta_j = \begin{cases} 2, & \text{if the review is made by a Vine reviewer} \\ 1, & \text{otherwise} \end{cases}$$

$$s_j = \begin{cases} \eta & \text{if matched words have a maximum grade of } \eta \\ 0, & \text{otherwise} \end{cases}$$

$$v_j = \begin{cases} 1, & \text{if the reviewer is a verified buyer} \\ 0.1, & \text{otherwise} \end{cases}$$

$$h_j = \begin{cases} e^{\frac{helpful\_votes_j}{total\_votes_j} - 0.5}, & \text{if } total\_votes_j > 0 \\ 1, & \text{otherwise} \end{cases}$$

$$L_j = \frac{1}{2}\log_{10} len_j$$

sj = 0时，Φ=0，所以加入修正项

$$\phi_{i,k,j}^{(t)} = \theta_j \cdot s_j^{v_j} \cdot e^{\frac{helpful\_votes_j}{total\_votes_j}-0.5} \cdot \frac{1}{2}\log_{10} len_j + \epsilon_j$$

$$\epsilon_j = \begin{cases} 1, & \text{if } s_j = 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\boldsymbol{A} = \text{diag}(a_{i,k,1}^{(t)}, a_{i,k,2}^{(t)}, \ldots, a_{i,k,j}^{(t)}, \ldots, a_{i,k,n}^{(t)})$$

根据评论中出现的情感词来决定±

$$a_{i,k,j}^{(t)} = \begin{cases} +1, & \text{if the corresponding review suggests a positive sentiment} \\ -1, & \text{if the corresponding review suggests a negative sentiment} \end{cases}$$

b. the Reputation Model(时间序列模型)
   i. Time Weight Sequence

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 & \gamma_2 & \ldots & \gamma_m & \ldots & \gamma_t \end{pmatrix}$$

$$\gamma_j = \frac{a^j}{a^t}, \quad a > 1$$

   a=1.1(常数)
   ii. Reputation

$$\boldsymbol{score} = \begin{pmatrix} score_{i,k}^{(1)} & score_{i,k}^{(2)} & \ldots & score_{i,k}^{(m)} & \ldots & score_{i,k}^{(t)} \end{pmatrix}^T$$

$$Rep_{i,k}^{(t)} = \boldsymbol{\gamma} \cdot \boldsymbol{score}$$

   iii. 模型评估：
   Trend Similarity between Quantified Reputation and Sales(作图(可视化)+Kendall's Tau Method);

c. the Successfulness Prediction Model(预测模型)
   i. Gaussian Process Regression
   https://www.zhihu.com/question/46631426?sort=created
   ii. 归一化，把(-∞，+∞）映射到[0, 1]

$$p_{i,k}(t) = \frac{1}{1 + e^{-\hat{Rep}_{i,k}^{(t)}}}$$

   iii. 判断成功的门槛：(早期产品有更大容错率)

$$threshold_{i,k}^{(t)} = 0.5 - 0.1e^{-\frac{1}{10}(t-\tau)}$$

5. results:
    a. rating和随之而来的review的关系：review使用第一个模型中的β，k个正/负面rating之后出现±review的频率加权和；
    b. rating和特定的评论词的关系：一些特定的sentiment words中的评分比例，出现在好/差评中频率高的词都是±面词；
6. strategy
7. Sensitivity Analysis：改变参数，观察某个特定值的变化；
8. Strengths and Weaknesses；
9. Conclusion ；
10. Letter；
11. References；