基于 EM 方法的销量异常识别

爱俊 <Aijun Bai, aijunbai@gmail.com>

2015 年 3 月 6 号

背景

商品的销量数据，经常有一些日销量非常高的情况，其中一些有作弊的嫌疑(比如 http://item.taobao.com/item.htm?spm=a1z10.3.w4002-289068221.13.fNsp1U&id=42751855145)，另外一些则看似正常（比如 http://detail.tmall.com/item.htm?spm=a230r.1.14.4.5ZZLqd&id=41682469998&ad_id=&am_id=&cm_id=140105335569ed55e27b&pm_id=&abbucket=19)。

本文的想法是使用隐马尔科夫模型来识别商品的销量数据属于作弊情况还是正常情况。假设任意商品有两种隐藏状态，分别是 Cheating 和 No-Cheating。Cheating 条件下，每天的销量数据有未知概率分布 $\Pr(Sales|\text{Cheating})$；No-Cheating 条件下，每天的销量数据有未知概率分布 $\Pr(Sales|\text{No-Cheating})$。

假设初始的观察模型、转移模型和概率分布。以该商品最近 N 天的销量数据作为该未知马尔科夫链的观察数据，通过 Expectation−Maximization (EM) 算法 [1] 估计出该未知马尔科夫链参数，进一步通过 Viterbi 算法 [2] 估计出最有可能的状态转移路径。

举例来说，输入的数据是最近 5 天的销量: 1000, 1200, 800, 6000, 5000。那么算法的输出可能是: No-Cheating，No-Cheating，No-Cheating，Cheating，Cheating 需要注意的是，已有的累积销量数据对买家也有影响（马太效应)，完整的模型需要考虑这个因素。另外，不同行业的初始概率分布也是不一样的。需要注意的是，这个模型的观察空间很大，是否实际可行，还需要仔细讨论。

方法

实现了使用 EM 算法来检测商品日销量数据变化异常的情况。假设商品最近 N 天的日销量数据为: $x_1, x_2, \ldots, x_N$ 假设正常情况下日销量服从参数为 λ_1 的泊松分布，作弊情况下服从参数为 λ_2 的泊松分布。$\lambda_2 > \lambda_1$，

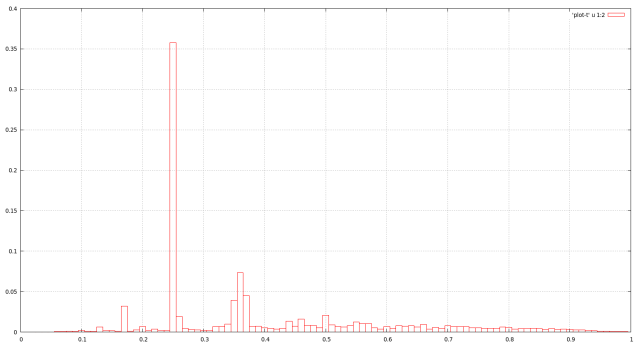

图 1: 作弊概率 p 的分布

λ_1、λ_2 未知。假设第 i 天销量数据 x_i 为作弊情况下的销量的概率为 p_i。$p_1, p_2, \ldots, p_N$ 未知。

注意到，已知 λ_1, λ_2 的情况下，可以使用贝叶斯方法计算出 $p_1, p_2, \ldots, p_N$ 的值；已知 $p_1, p_2, \ldots, p_N$ 的情况下，可以使用极大似然估计计算出 λ_1, λ_2 的值。

所以可以使用 EM 的方法，假设 λ_1, λ_2 的初始值，估计 $p_1, p_2, \ldots, p_N$ 的值，进一步估计 λ_1, λ_2 的值，重新估计 $p_1, p_2, \ldots, p_N$ 的值，直到收敛，估计出最有可能的 $p_1, p_2, \ldots, p_N$ 以及 λ_1, λ_2 的值。

最后，认为该销量数据 $x_1, x_2, \ldots, x_N$ 有 $\frac{\max(\lambda_1,\lambda_2)-\min(\lambda_1,\lambda_2)}{\max(\lambda_1,\lambda_2)}$ 的概率（记为 p）存在作弊行为，这个概率解释还可以继续细化。

实验验证

使用 2014.12.14 到 2014.12.18 五天的数据，检测出来的商品日销量存在波动异常数据的作弊概率 p 分布和累积分布分别如图 1和 2。举例来说，逗妮开心旗舰店的商品 39284943258（逗妮开心巴旦木坚果零食特产手剥巴旦木 218 克 x3 袋）这五天的销量数据分别为：0, 0, 1, 17466, 30649，算法输出作弊的概率是：0.999967；anmashu 旗舰店的商品 40237454086（2014 冬季新款韩版保暖裤高端双面羽绒裤小脚裤白鸭绒长裤女靴裤）这五天的销量数据分别为：9, 13, 1, 1, 3498，算法输出的作弊概率为：0.999714。

一些定量结果：

- 作弊概率 $p > 0.9$ 的商品占了商品总数的 1.5%
- 作弊概率 $p > 0.75$ 的商品占了商品总数的 8.1%
- 作弊概率 $p > 0.5$ 的商品占了商品总数的 26.9%

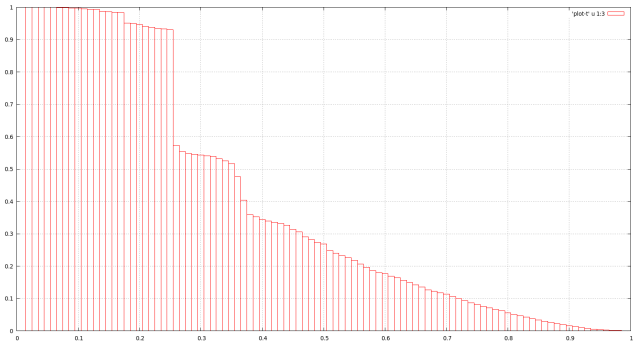

图 2: 作弊概率 p 的累积分布

下一步，需要进一步做好作弊概率 p 的解释，同时还需要处理销量数据不足 N 天的商品的作弊概率计算和比较。

代码

截至本文档完成之日，完整的代码位于团队 SVN 目录：http://svn.develop.taobao.net/repos/searchAppCodes/AntiSpam_Tsc/trunk/aijun.baj/abnormal_sales_detection/。本节主要介绍现有代码的主要结构、实现细节和运行示例。

目录结构

目前的主要目录结构如下：

conf/: 一些配置文件

doc/: 相关文档

py/: Python 源文件

analysis.sh: 数据分析脚本

plot.sh: 可视化脚本

run.sh: 处理数据并分析脚本

test-offline.sh: 使用离线构造数据测试脚本

test-online.sh: 使用在线数据测试脚本

实现细节

一些值得指出的实现细节如下:

- Decimal 的使用是为了实现高精度浮点数计算。
- EM 算法的最大迭代次数设为 100。
- EM 算法的收敛条件是相邻更新的距离小于 0.001。

运行示例

一些脚本调用示例如下:

./run.sh -n 10: 分析最近 10 天的销量数据，并可视化。

./test-offline.sh: 构造数据，并离线测试 Map-Reduce 流程

./test-online.sh: 使用在线数据测试 Map-Reduce 流程

总结

本文提出基于 EM 算法的销量异常检测方法，实际数据上的实验结果表明算法可以检测出来销量波动异常的情况，但波动异常背后的真实原因还需要进一步分析。

参考文献

[1] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39, 1–38 (1977)

[2] Viterbi, A.J.: Viterbi algorithm. Scholarpedia 4(1), 6246 (2009), http://www.scholarpedia.org/article/Viterbi_algorithm