**项目报告·数据挖掘作业二：关联规则挖掘**

梁天行 5720182040

**数据源**：oakland-crime-statistics-2011-to-2016

**数据预处理**

对数据集进行处理，转换成适合关联规则挖掘的形式：

由于数据集过于庞大，此次实验仅手动选取2016一年的数据；

源纪录中包括的键值有“Agency、Create Time、Location、Area Id、Beat、Priority、Incident Type Id、Incident Type Description、Event Number、Closed Time”，此次实验以探究不同犯罪案件类型间的潜在联系为目的，用Apriori算法，统计不同地区犯罪率最高的案件。因此对源数据进行处理，只保留键值为“Area Id”、“Incident Type Id”的数据。以上过程主要由代码中 “Partition” 、“Likely\_Crime” 、“LoadDataSet”函数完成。

**找出频繁项集，导出关联规则，计算其支持度和置信度**

运用Apriori算法，首先会生成所有地区犯罪率最高的案件种类项集列表；

扫描案件种类记录记录来查看哪些项集满足最小支持度要求，那些不满足最小支持度的集合会被去掉；

对剩下的集合进行组合以生成包含两个元素的项集；

接下来重新扫描交易记录，去掉不满足最小支持度的项集，重复进行直到所有项集都被去掉。（人为定义最小支持度，支持度：一个项集的支持度被定义为数据集中包含该项集的记录占总记录的比例。）此过程主要由代码中 “ScanD”函数完成。

置信度就是根据某一个条件，得到一个结论的可信程度、可靠程度。本实验中采取的计算方式是两条案件记录的支持度之比。此过程主要由代码中 “calcConf” 、“rulesFromConseq” 、“generateRules”函数完成。

对结果的评测采用lift和全自信度，依据以下公式进行计算：

lift(X→Y)=lift(Y→X)=conf(X→Y)supp(Y)=conf(Y→X)/supp(X)=P(X⋂Y)/P(X)P(Y)

all\_confidence(A,B)=P(A⋂B)/max{P(A),P(B)}

=min{P(B|A),P(A|B)}

=min{confidence(A—>B),confidence(B—>A)}

**分析结果**

在规定统计区域犯罪率最高案件数位为15，最小支持度0.5，最小置信度为0.7时，得到以下结果：见out.txt

对结果简单分析可知，在各地区犯罪率高的案件都是频繁发生的，由此得到的分析结果显示出各种不同案件中有较高的关联程度。同时基于测试集规模的选择，选取的犯罪案件种类有限，以及分地区作为不同记录的机制仍有局限性，因此只能发掘出最浅显的以案件发生次数为依据的关联关系。