课程设计报告

软件工程

软1207

车晓宇 201292100

任强 201292148

课程设计日期：2015年7月20日

# 问题说明

Top-K条件判别模式挖掘实际上是一个关联分析的子集，也就是说，Top-K条件判别模式挖掘是关联分析的一种。关联分析是一种用于在大型数据集中发现数据之间有意义的联系，从而指导人们的行为。而通过关联分析而发现的联系可以用关联规则或频繁项集的形式来表示。

具体来说，在这个问题中，每条实验数据都代表了一个样本，而每个样本都有着各自的属性。除了属性以外，所有样本都有一个class分类属性，相同的class属性的样本属于同一类。

而模式由属性和属性值组合而成，不同模式一定有着不完全相同的属性或者属性值。前景集合是其中某一类数据的集合，而背景集合是其余类别的数据构成的集合。而频率是指某个数据集合包含某个模式的实例数据的百分比，也就是满足模式的样本集占某类集合的百分比。

全局差异度指某模式在前景集合和背景集合频率的差值。而局部差异度涉及了子模式，而子模式实际就是模式的子集。局部差异度由给出的公式即可求出。

而要解决的问题，则是要通过设计一个快速算法来找到Top-K的模式，分别求前k个全局差异度最大的模式和前k个局部差异度最大的模式。

# 算法描述

在频繁模式挖掘算法中，目前已经存在三种经典算法：Apriori算法，FP-growth算法和基于垂直数据格式的算法。基于这三种种基本的经典算法，又可以通过一些改进来得到他们的改良算法。

首先要处理得到的数据集。在从文件中读到Python的二维list中之后，便可以反复对该list的数据进行读取。其中包括数据集的各个属性及对应的属性值，最后一列一般为class，用于分类前景集合及后景集合。

利用Python中的dict结构来存储模式pattern。每个pattern都包括