本文档对诊疗过程异常发现部分的程序进行介绍。首先对程序的交互界面进行介绍，然后从以下五部分对“诊疗过程异常发现”这一功能进行介绍：数据导入、数据预处理、基于 LDA 的诊疗主题聚类、基于 K-means 的天事件聚类和基于改进的成本函数的诊疗日志合规性检查等。

1. 交互界面介绍

本程序的交互是基于 JavaFX 构建的，程序入口为 view.FrameworkMain.java，初始界面如图 1.1 所示。区域划分如图 1.2 所示，视图选择区有两个视图可供选择，其中数据视图用于查看所导入的原始数据，功能视图包含在原始数据上执行的各种功能性处理操作，比如“临床路径挖掘”和本文档的“诊疗过程异常发现”等。菜单栏提供相应视图下各功能的入口。结果展示区是一系列 Tab 的集合，用于展示数据以及数据处理的结果。

图 1.1　初始界面

图 1.2　初始界面的区域划分

2. 数据导入

本小节以“文件导入方式”为例对数据的格式要求以及数据导入部分的功能

进行介绍。文件格式为 csv 格式，由逗号分隔，各列标题为：visitId（就诊标识），event（诊疗项目标识），eventClass（诊疗项目类别），num（数量），price（单价），time（日期），diagnoseId（诊断编码），hospitalId（医疗机构编码），departmentId（科室编码）和 doctorId（医生编码）等，其中 visitId、event、num 和 time 为必选项。

具体的导入流程为：首先选择“数据视图”，然后选择菜单栏中的“导入”菜单，最后选择其中的“从文件导入”。导入之后界面如图 2.1 所示，可以对所导入的数据进行“删除”或者“查看”。点击“查看”之后的界面如图 2.2 和图 2.3 所示，其中前者显示患者数量、诊疗项目数量和住院天数的分布等统计结果，后者显示数据详情（包含患者编号，诊疗项目，项目类别，数量，单价和日期等）。

图 2.1 数据集展示

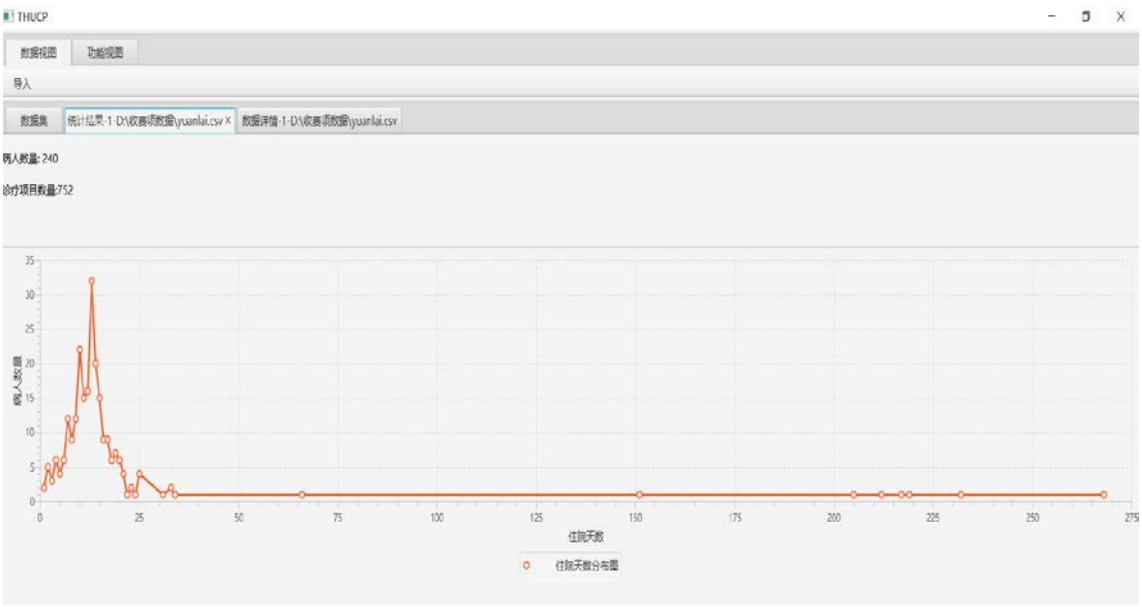


图 2.2 住院天数分布结果展示

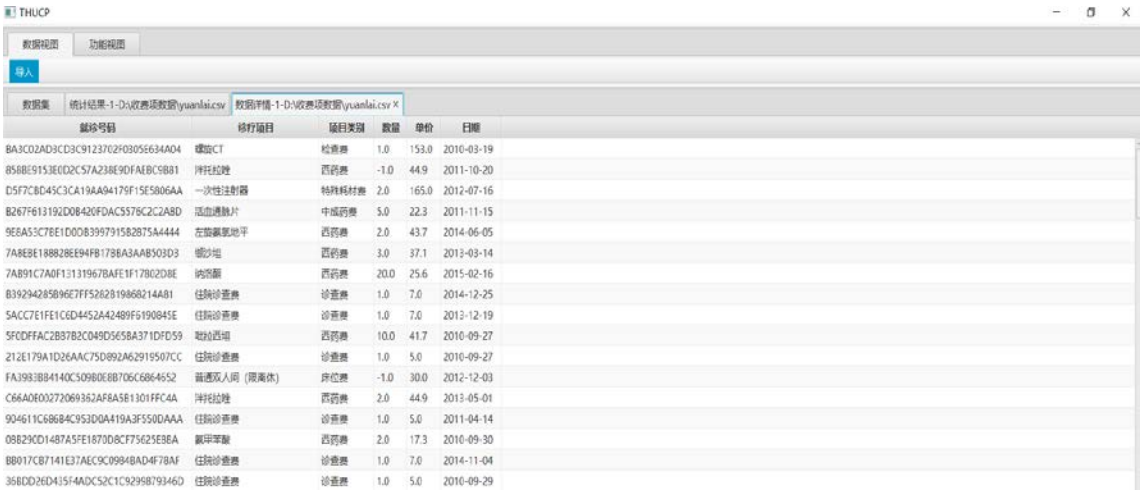


图 2.3 数据详情展示

3. 数据预处理

数据预处理的使用过程为：首先选择“功能视图”，然后选择菜单栏中的“诊疗过程异常发现”，然后选择“数据预处理”。在开始之前需要建立或者选择工作空间，用于隔离不同数据集以及在其之上的处理操作所产生的结果，如图 3.1 所示。选择工作空间之后，选择所要处理的临床数据（或者收费项数据），即选择之前导入的数据，如图 3.2 所示。经过以上设定之后预处理部分便开始执行，处理之后的结果文件所在路径为：workspace/LDA/doc2items.csv、workspace/LDA/id2item.csv。其中的 workspace 是所选择的工作空间的路径。doc2items.csv 是满足 LDA 输入格式的数据，每一行的形式为“B772F6#2014-10-10,-,317 317 315 315…”，代表某一天的诊疗项目， 第一列的“B772F6#2014-10-10”为天标识（其中 B772F6 为病人标识，2014-10-10 为当天的日期），第二列用于验证分类效果，在本程序中本列无用，第三列为空格分隔的诊疗项目标识（标识与诊疗项目名称的对应关系在 id2item.csv 这一结果文件中给出）。经过以上预处理操作之后，便可以应用 LDA 算法得到每天的诊疗主题分布，并将诊疗项目按照诊疗主题进行聚类。具体的预处理细节可以参见论文的 4.1 小节。

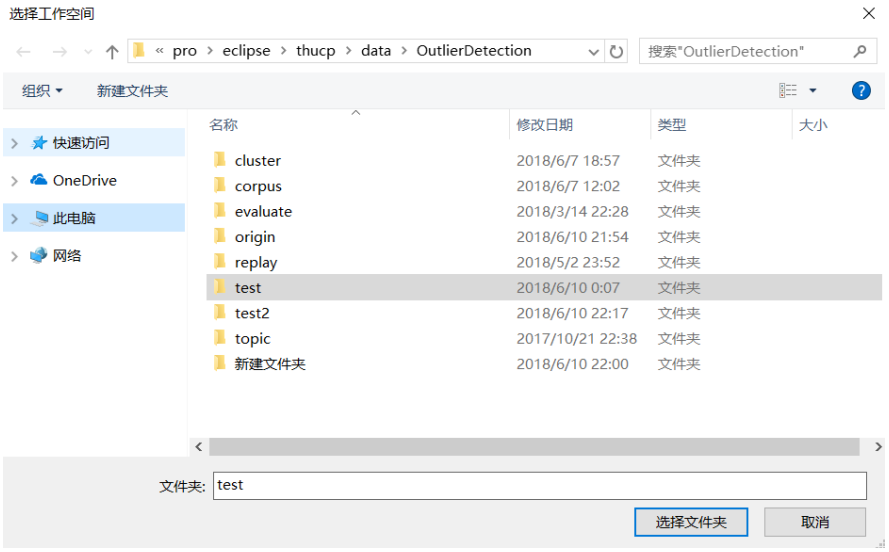


图 3.1 工作空间选择界面

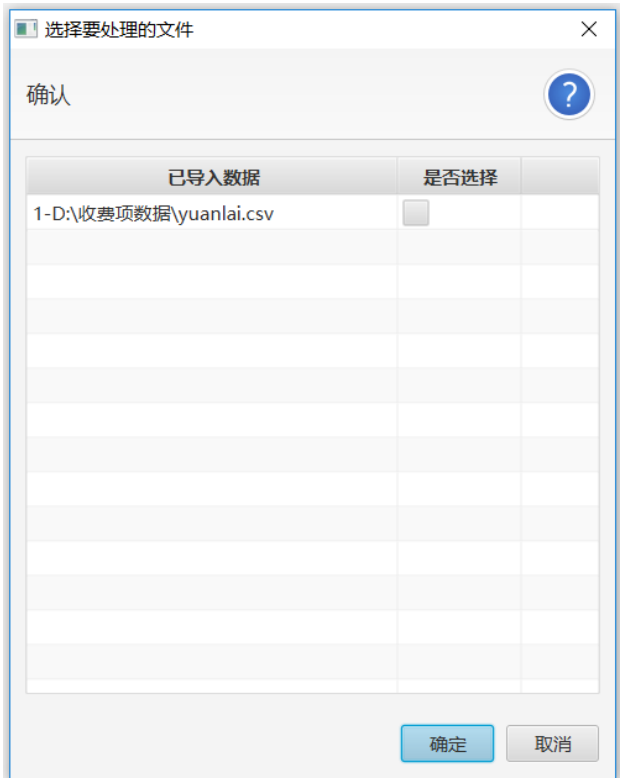


图 3.2 文件选择界面

4. 基于LDA的诊疗主题聚类

这一部分的使用过程为：首先选择“功能视图”，然后选择菜单栏中的“诊疗过程异常发现”，然后选择“基于 LDA 的诊疗主题聚类”。同样，在开始之前需要选择工作空间，然后设置 LDA 相关的参数，如图 4.1 所示。在运行 LDA 算法之前需要确定 K 的取值，本文是基于困惑度随 K 值的变化曲线的拐点来进行 K 值

选取的。LDA 的 K 值起点和终点确定了 LDA 的 K 值的探索范围，迭代次数为 LDA 算法的迭代次数，重复次数是指对于每个 K 值，LDA 算法的运行次数（由于 LDA 结果具有一定的随机性，所以需要运行多次，然后从中选择主题聚类效果较好的结果）。LDA 结束之后会得到困惑度随 K 值的变化曲线，如图 4.2 所示，该曲线仅供参考，可以根据需要选择其他专业的统计软件来绘制散点图或者拟合后的曲线图（困惑度和 K 值的对应结果的路径为 workspace/LDA/evaluate.csv，可以将此数据导入到其他专业统计软件中）。LDA 主题聚类的结果文件为 workspace/LDA/doc2topics-3-0.csv 和 workspace/LDA/topic2items-3-0.csv（以 K=3 为例），其中 doc2topics-3-0.csv 记录了所有天事件的诊疗主题概率分布，topic2items-3-0.csv 记录了各诊疗主题下的所有诊疗项目的概率分布，具体原理参见论文的 2.2 小节；3 代表 K 的取值，0 代表该取值下的某次 LDA 主题聚类结果的序号（如果重复次数为 5，则会生成 5 个 LDA 主题聚类结果，序号为 0~4）。

根据困惑度随 K 值的变化曲线确定 K 值之后，便可以在设置图 4.1 所示的参数时，将起点和终点都设置为选定的 K 值，然后将重复次数设置为 100，最终在运行结束之后从 100 个结果中选择聚类效果最好的结果（选择原则可以参见论文的 6.2.1 小节）。

图 4.1 LDA 相关参数设置界面

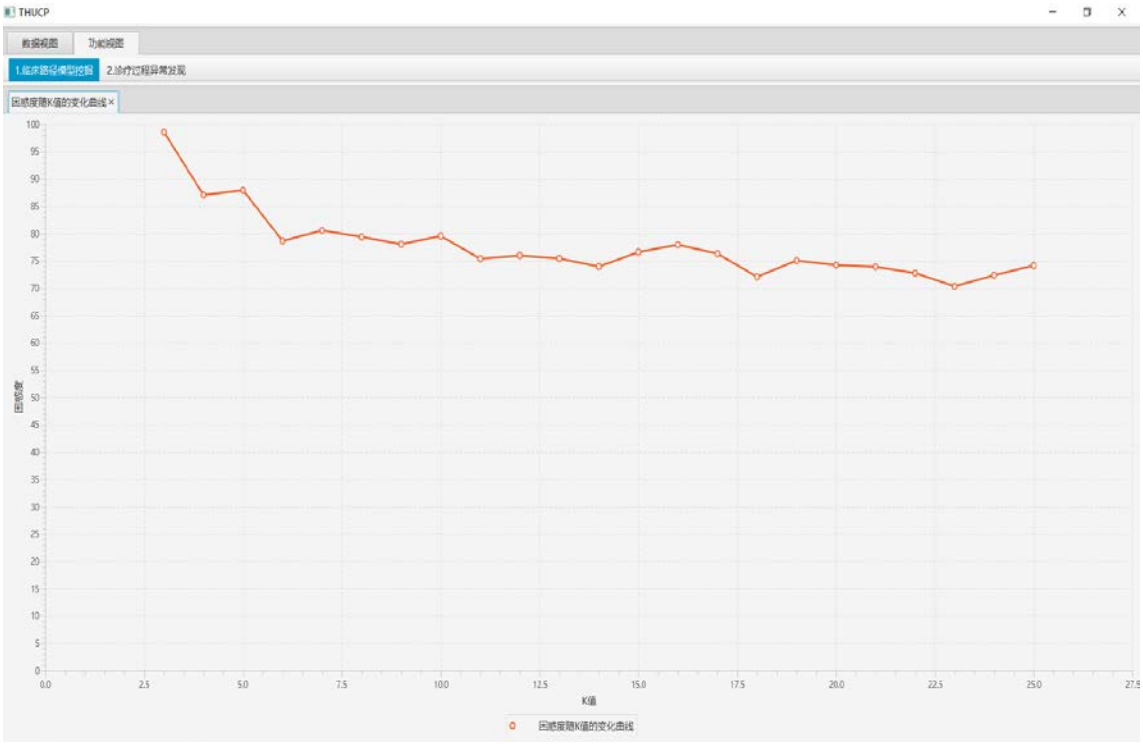


图 4.2 困惑度随 K 值的变化曲线

5. 基于K-means的天事件聚类

这一部分的使用过程为：首先选择“功能视图”，然后选择菜单栏中的“诊疗过程异常发现”，然后选择“基于 K-means 的天事件聚类”。同样，在开始之前需要选择工作空间，然后设置 K-means 相关的参数，如图 5.1 所示。其中，LDA 的 K 值和序号用于定位本文档的第 4 小节中得到的效果最好的 LDA 主题聚类结果，作为天事件聚类的输入。K-means 的 K 值的起点和终点确定了 K 值的探索范围，用于根据此范围内误差平方和随 K 值的变化曲线得到 K 的最终取值。运行结束之后，同样给出如图 5.2 所示的曲线以供参考（具体的坐标结果可以参见 workspace/K-means/evaluate.csv）。另外，给出了每个 K 值所对应的诊疗过程模型挖掘结果，图 5.3 展示的是 LDA 的 K 值选取 16、序号选取 16、K-means 的 K 值选取 11 的结果（经过天事件聚类之后的日志文件的路径为：workspace/K-means/kmeans-11.csv）。图 5.4 展示的是每个类簇的代表性诊疗主题（结果文件的路径为 workspace/K-means/cluster2topics.csv），比如图中显示的 cluster-0 这一类簇的代表性诊疗主题为 topic-14、topic-4 和 topic-11。另外，对于每个 topic，给出了其代表性诊疗项目。

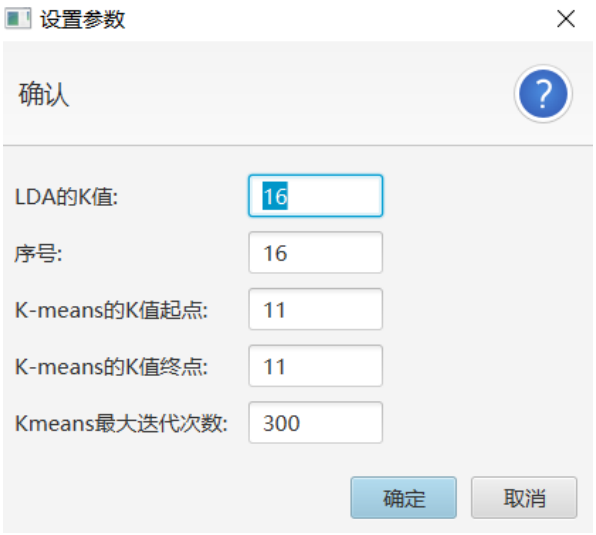


图 5.1 K-means 相关参数设置

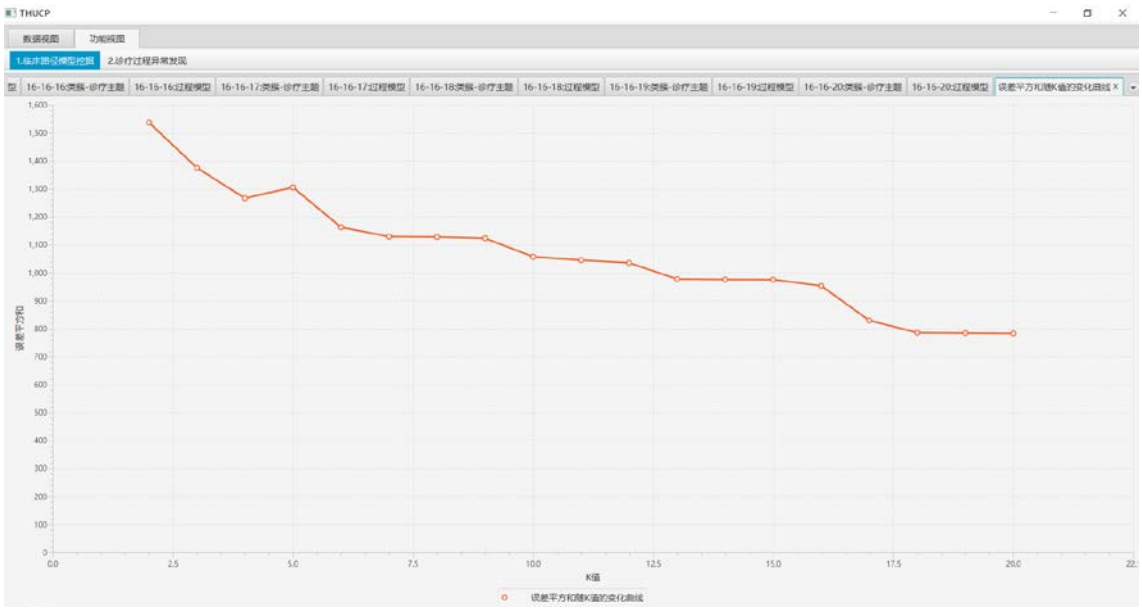


图 5.2 误差平方和随 K 值的变化曲线

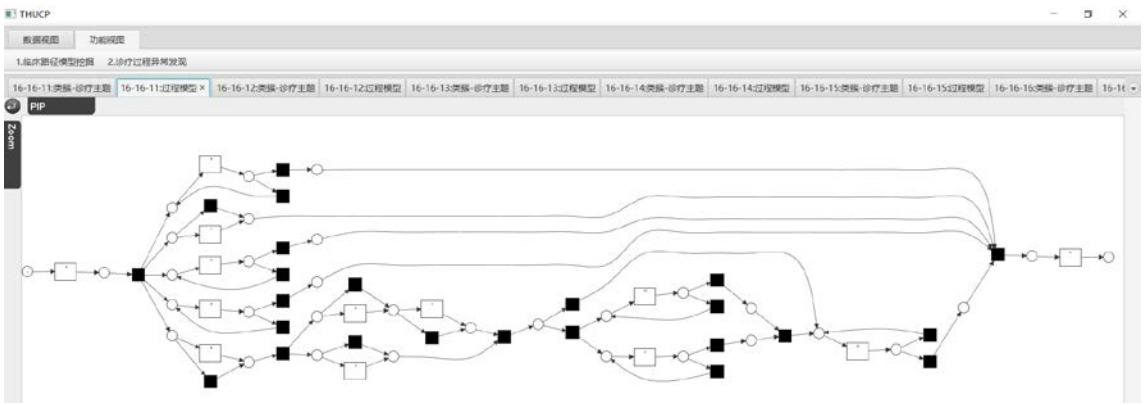
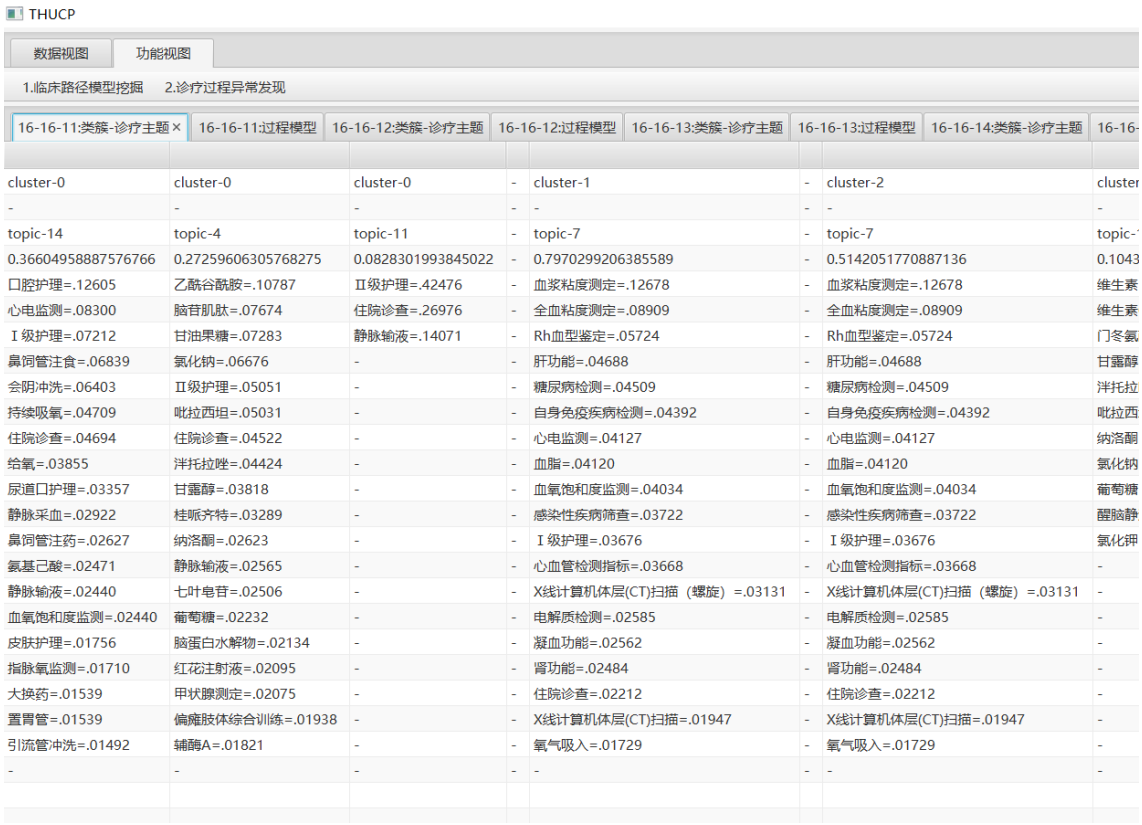


图 5.3 诊疗过程模型



cluster-0	cluster-0	cluster-0	-	cluster-1	-	cluster-2	cluster
-	-	-	-	-	-	-	-
topic-14	topic-4	topic-11	-	topic-7	-	topic-7	topic-
0.36604958887576766	0.27259606305768275	0.0828301993845022	-	0.7970299206385589	-	0.5142051770887136	0.1043
口腔护理=.12605	乙酰谷酰胺=.10787	Ⅱ级护理=.42476	-	血浆粘度测定=.12678	-	血浆粘度测定=.12678	维生素
心电监测=.08300	脑苷肌肽=.07674	住院诊查=.26976	-	全血粘度测定=.08909	-	全血粘度测定=.08909	维生素
Ⅰ级护理=.07212	甘油果糖=.07283	静脉输液=.14071	-	Rh血型鉴定=.05724	-	Rh血型鉴定=.05724	门冬氨
鼻饲管注食=.06839	氯化钠=.06676	-	-	肝功能=.04688	-	肝功能=.04688	甘露醇
会阴冲洗=.06403	Ⅱ级护理=.05051	-	-	糖尿病检测=.04509	-	糖尿病检测=.04509	泮托拉
持续吸氧=.04709	吡拉西坦=.05031	-	-	自身免疫疾病检测=.04392	-	自身免疫疾病检测=.04392	吡拉西
住院诊查=.04694	住院诊查=.04522	-	-	心电监测=.04127	-	心电监测=.04127	纳洛酮
给氧=.03855	泮托拉唑=.04424	-	-	血脂=.04120	-	血脂=.04120	氯化钠
尿道口护理=.03357	甘露醇=.03818	-	-	血氧饱和度监测=.04034	-	血氧饱和度监测=.04034	葡萄糖
静脉采血=.02922	桂哌齐特=.03289	-	-	感染性疾病筛查=.03722	-	感染性疾病筛查=.03722	醒脑静
鼻饲管注药=.02627	纳洛酮=.02623	-	-	Ⅰ级护理=.03676	-	Ⅰ级护理=.03676	氯化钾
氨基己酸=.02471	静脉输液=.02565	-	-	心血管检测指标=.03668	-	心血管检测指标=.03668	-
静脉输液=.02440	七叶皂苷=.02506	-	-	X线计算机体层(CT)扫描（螺旋）=.03131	-	X线计算机体层(CT)扫描（螺旋）=.03131	-
血氧饱和度监测=.02440	葡萄糖=.02232	-	-	电解质检测=.02585	-	电解质检测=.02585	-
皮肤护理=.01756	脑蛋白水解物=.02134	-	-	凝血功能=.02562	-	凝血功能=.02562	-
指脉氧监测=.01710	红花注射液=.02095	-	-	肾功能=.02484	-	肾功能=.02484	-
大换药=.01539	甲状腺测定=.02075	-	-	住院诊查=.02212	-	住院诊查=.02212	-
置胃管=.01539	偏瘫肢体综合训练=.01938	-	-	X线计算机体层(CT)扫描=.01947	-	X线计算机体层(CT)扫描=.01947	-
引流管冲洗=.01492	辅酶A=.01821	-	-	氧气吸入=.01729	-	氧气吸入=.01729	-
-	-	-	-	-	-	-	-

图 5.4 各类簇的代表性诊疗主题

6. 基于改进的成本函数的诊疗日志合规性检查

这一部分的使用过程为：首先选择“功能视图”，然后选择菜单栏中的“诊疗过程异常发现”，然后选择“基于改进的成本函数的诊疗日志合规性检查”。同样，在开始之前需要选择工作空间，然后设置 K-means 的 K 值（即在本文档的第 5 小节中所选定的 K 值，用于定位相应的日志文件用于本小节的合规性检查），如图 6.1 所示。运行结束之后会显示从日志文件中得到的诊疗过程模型（如图 6.2 所示）以及异常发现结果（如图 6.3 所示）。对齐结果（或者合规性检查结果）的路径为：workspace/replay/alignment.csv。

图 6.1 合规性检查相关的参数设置

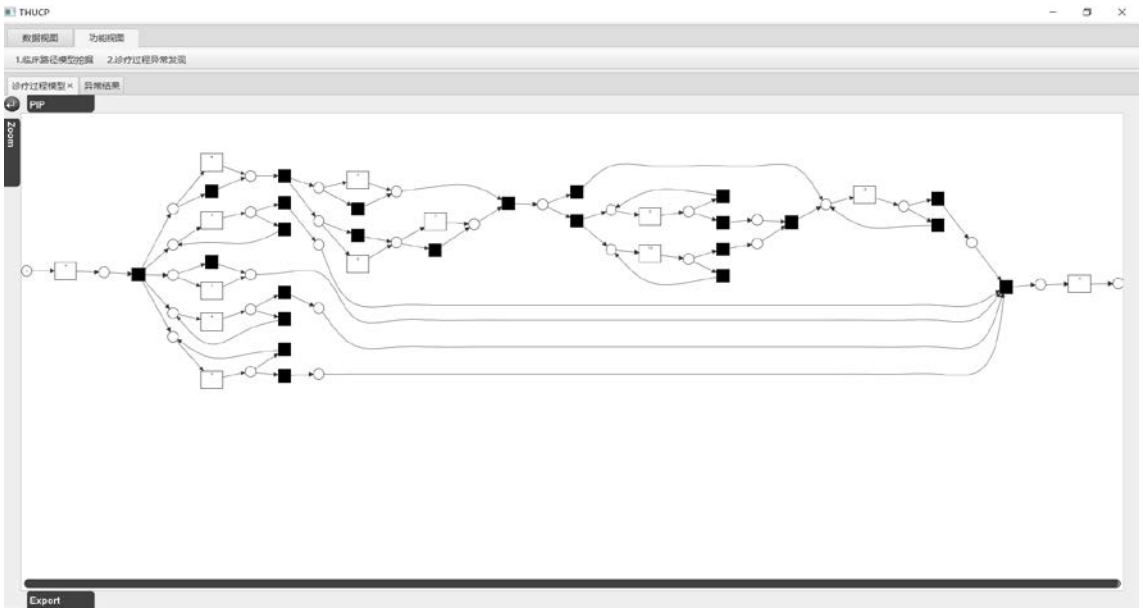


图 6.2 诊疗过程模型

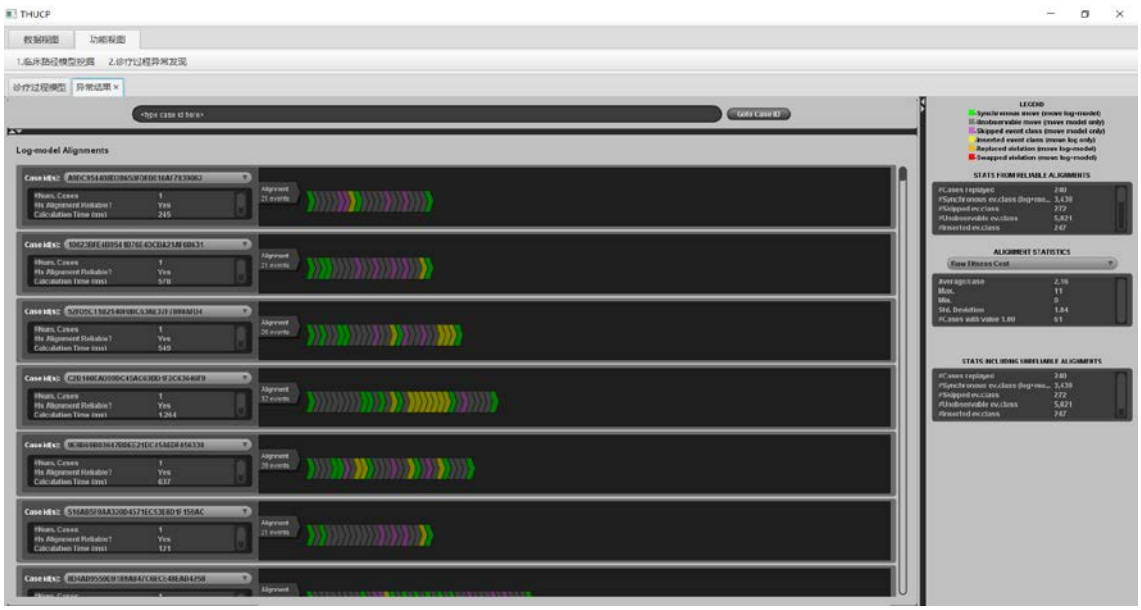


图 6.3 合规性检查结果