Trace Cluster by Using Frequent Sequence Patterns

Project Initiation

**1 Overview**

The emergence of process discovery in many fields, led to large-scaled data collection. Especially in the health care sector large amounts of data are collected. This data however is usually a heterogenous collection of all patients, while many process mining algorithms need a homogenous data set to work correctly.

In order to obtain homogenous data sets out of the variety of data collected, only two possibilities exist. Either experts on the field cluster the data by hand, which is not feasible due to the amount of data collected, or algorithms are used which try to cluster the data based on metrics and parameters.

One method to obtain such clusters is to apply clustering based on sequence patterns. Here the single traces in the data are tested for theoccurence of certain patterns and traces exhibiting the same patterns are assigned to the same cluster.

Oftentimes however the clusters generated by algorithms do not match the clusters created by experts, due to a difference in weighting of patterns between algorithms and real life use case or entirely different patterns being applied.

This project aims to connect both approaches, by obtaining the clustering criteria used by the experts with the scalability of the clustering algorithms, by creating an algorithm in python allowing an expert to create clusters on a small sample set and then analyzing these clusters for criteria to aplly to the whole data set.

**2 Business Case**

In this section we discuss the business to describe the process further. First we outline the use case on the example of health care data. Furthermore we describe the scope oft he project and point out the key benefits.

**2.1 Use case**

As a use case we present the situation in the health care sector. A large amount of patient data is collected every year. Since this data is mostly collected at central stations such as the hospitals and health insurances, the data is heterogenous. Data for all kinds of patients is collected together with different diseases, where single patients can have multiple diseases as well. In order to apply and profit from process mining algorithms, the data needs to be clustered according to the diseases. While medical expert can assign patient traces to correct clusters by analysing the treatment entries stored in the traces, learning which criteria are important for certain clusters is too complex for an algorithm to perform automatically. Thus giving the algorithm input of a medical expert which criteria are important drastically increases the clustering. Furthermore this project allows the medical expert to give the criteria implicit by simply clustering a small set of samples by hand so that the algorithm can determine the criteria out of the sample set.

**2.2 Scope**

During the project both the algorithm and a user interface will be developed, thus the scope is divided into these two subprojects:

**The algorithm:**

* Mine FSPs from a given small sample set of traces using sequence mining
* Use FSPs to score all traces in the event log
* Filter traces by thresholds
* Calculate the cluster’s quality measurements
* (determine thresholds automatically)

**The user interface:**

* Select (and label) a small set of traces
* Set the threshold
* List FSPs for groups
* Show quality measurements
* Set thresholds fort he score generated by FSPs for selecting traces

Besides the code, also the following documentation will be provided:

* Project Initiation Document
* Requirements analysis

**2.3 Key Benefits**

Working on homogenous data can improve the results of many process mining algorithms. Generating homogenous clusters from real world data, that match actual real world criteria is a difficult task, leading to incorrect clusterings and thus overall bad performance of process discovery in these cases. With this project we aim to create clusters that match the real world criteria well even for large data sets.

Additionally creating a user interface to select the correct clusters in the sample allows medical experts without knowledge in software, to perform the clustering and thus provide good criteria for the clustering oft he complete data set.

**3 Feasibility Study**

In this section we analyze the feasibility of the described project. This is done from a theoretical and a technical point of view. In addition we identify potential risks and provide mitigation strategies.

**3.1 Theoretical Point of View**

From a theoretical point of view, there already exist pattern based clustering algorithms, such as the pCluster algorithm. The limitations of these algorithms lie mainly in the size of the potential data set, but also in determining the correct criteria to choose for a cluster.

In our project we focus on finding the weighting of patterns based on input of real life experts. This allows to generate more precise clusters, while at the same time limiting the patterns to be tested to length 1 and 2 and also testing closed sequences.

Since the algorithm has already been implemented in ProM yielding good results, we conclude that the project is feasible from a theoretical point of view.

**3.2 Technical Point of View**