Filière Informatique de gestion
Studiengang Wirtschaftsinformatik

Hes·so VALAIS WALLIS

Data mining in Bioprocesses – Knowledge discovery in databases (IIG part)

Student : Christopher Artero
Professor : René Schumann

Summary

1. Is it possible to annotate historical data by clustering them?
2. Determine the most suitable clustering algorithm for the current type of data
3. Implement the algorithm and a pattern matching with domain knowledge

Introduction

- **The Institute of life technologies** of the HES-SO VS gathered data from experimentations that come from **four bioreactors.**
- There is no annotation available for the data.
- They want to know if it is possible to annotate the data by clustering them.

Methods

- Retrieving data from the four old databases
- State-of-the-art of the existing clustering algorithms
- Defining the most suitable algorithm
- Implementation of a pattern matching sequence, 3 atomic patterns and 3 composed patterns
    - **Atomic**
        - Sterilisation
        - “Batch”
        - Chemostat
    - **Composed**
        - Batch
        - Fed-Batch
        - Chemostat
- Implementation of the clustering algorithm
- Comparison of the results
- Discussion with the people from the ITV

Results

- Pattern Matching (Greybox):
    - 8 Batch
    - 1 Chemostat
    - 10 Fed-batch
    - 12 Other

- K-Medoids 4 (Blackbox) with cluster named based on the results of the pattern matching sequence:

Cluster	Count(Exp)	Possible exp.name based on greybox results
Row0	1	(1) Chemostat
Row2	7	(8) Batch
Row13	12	(12) Other
Row23	11	(10) Fed-Batch

    - Percentage of experimentation in the right cluster based on the result of the greybox: **35%**

- K-Medoids 3 Clusters with cluster named based on the results of the pattern matching:

Cluster	Count(Exp)	Possible exp.name based on greybox results
Row13	10	Batch
Row23	5	Fed-Batch
Row28	4	Chemostat

    - Percentage of experimentation in the right cluster: **42%**

Conclusions

- By doing the **Pattern Matching** sequence, the results are good and can be qualified as the ground truth and allow the annotation of the data
- The **K-Medoids algorithm** produced bad results but there is multiple reasons for that
    - 31 experimentations for 4 clusters
    - 1 row to represent several days or weeks
    - A third of the data is clustered as “Other” based on the pattern matching results
- The clusters can be improved in the future
    - New patterns that are more accurate or anti-pattern
    - More data
    - Manage differently all failed or corrupted experimentation

Contact and more informations:
http://www.hevs.ch/ig

swissuniversities EQUAL SALARY Valais excellence CERTIFIED