Ghent University - TELIN - DDCM
2021-2022

Assignment 4: Categorical Edit Rules

1 Introduction

During this final assignment, you will test your skills related to categorical edit rules acquired during the corresponding theory lecture and lab session. In particular, you will be asked to solve exercises related to the concepts of categorical edit rules and to assess the consistency of a dataset persisting data about the type and purpose of clinical trials reported by both DRKS[1] and ClinicalTrials.gov[2]. In the following, some practical information is given in Section 2 and the exercises you have to solve are given in Section 3.

2 Practical information

For this assignment, make sure to take into account the following practical guidelines.

- The assignment has to be solved individually. If any form of irregularities are detected (e.g. plagiarism, collaboration,...), we will intervene in an appropriate way.
- To solve the exercises, you can use any of the provided (course) material (papers, online documentation, presentations, solved exercises,...) and any tool (PostgreSQL, pgAdmin, rulebox,...).

[1] `https://www.drks_web/`
[2] `https://www.clinicaltrials.gov/`

- For most of the exercises, you need the data that are persisted in the `purpose` table of the `clinicaltrials` database. You can connect to this database by using the following connection parameters.
  - Hostname/IP-address: ddcmstud.ugent.be
  - Port: 8088
  - Database: clinicaltrials
  - Username: sql_exerciser
  - Password: 7UCVuJeLCGcQbk2M[3]

  You can only connect with the database if you are within the UGent network (or by means of a VPN-connection to the UGent network). In Appendix 1, you can find an overview of all attributes and their values that appear in this table. For the attributes, we have provided a short explanation of their meaning.

- To submit your solutions, we expect you to upload one .zip file with the name `studentcode_firstname_lastname_assignment4.zip`. In this filename, you obviously have to substitute 'studentcode', 'firstname' and 'lastname' by your student code, first- and lastname (without spaces) respectively. For most of the exercises, we expect you to write a *short* answer in a written report (in any format of your choice). For some of the exercises, we expect you to create a specific file (e.g. .rbx file,...). We will clearly indicate, for each exercise, what we are expecting to receive, but in the end, all files should be added to the .zip file. Submitting can be done via the 'Assignments' module of Ufora and we will only look into the last submission uploaded before the deadline.

- The deadline to submit this assignment is May 20th, 2022, 10:00 PM.

- This assignment is part of the non-periodic evaluation (NPE). The NPE will account for 70% of the total score.

- To pass this course, you should obtain a minimum score of 10/20 for the NPE. Partial exemption for the NPE (retake) is possible if you have passed this part in the first evaluation period. If not, you should complete a modified version of the entire NPE during the summer.

- Unfortunately, If any of the stated requirements are not met, you will lose all points for this assignment. This is, for example, the case if filenames are not correct, files are not executable or unreadable, solutions are submitted after the deadline,... Therefore, make sure to check everything before submitting. If any form of irregularities are detected, further steps will follow.

- If you have any additional questions or issues, do not hesitate to contact us.

[3] Do not copy the password from this file, but insert it manually to avoid any mistakes.

3 Exercises

In order to assess the consistency of the data in the `purpose` table, we have provided a set of 17 edit rules, denoted as $\mathcal{E}$ and listed below. Attributes that are not involved in a rule, are discarded. Before starting with the exercises, make sure to read carefully through the set of edit rules and try to understand the underlying semantics of the rules.

- E^1: `stni`:{Epidemiological study, Observational study, Other}× `cp`:{Treatment}
- E^2: `stni`:{Epidemiological study, Observational study, Other} × `dp`:{Treatment}
- E^3: `cp`:{Basic Science, Diagnostic, Health Services Research, Other, Prevention, Supportive Care, Treatment} × `dp`:{Screening}
- E^4: `cp`:{Basic Science, Diagnostic, Health Services Research, Other, Prevention, Supportive Care, Screening} × `dp`:{Treatment}
- E^5: `cp`:{Basic Science, Health Services Research, Other, Prevention, Supportive Care, Screening, Treatment} × `dp`:{Diagnostic}
- E^6: `cp`:{Diagnostic, Health Services Research, Other, Prevention, Supportive Care, Screening, Treatment} × `dp`:{Basic research/physiological study}
- E^7: `cp`:{Basic Science, Diagnostic, Health Services Research, Other, Prevention, Screening, Treatment} × `dp`:{Supportive care}
- E^8: `cp`:{Basic Science, Diagnostic, Other, Prevention, Supportive Care, Screening, Treatment} × `dp`:{Health care system}
- E^9: `cp`:{Basic Science, Diagnostic, Health Services Research, Prevention, Supportive Care, Screening, Treatment} × `dp`:{Other}
- E^{10}: `cp`:{Basic Science, Diagnostic, Health Services Research, Prevention, Supportive Care, Screening, Treatment} × `dp`:{Prognosis}
- E^{11}: `cp`:{Basic Science, Diagnostic, Health Services Research, Other, Supportive Care, Screening, Treatment} × `dp`:{Prevention}
- E^{12}: `cst`:{Interventional} × `stni`:{Epidemiological study, Observational study, Other}
- E^{13}: `cst`:{Observational, Observational [Patient Registry]} × `stni`:{N/A}
- E^{14}: `dst`:{Interventional} × `stni`:{Epidemiological study, Observational study, Other}
- E^{15}: `dst`:{Non-interventional} × `stni`:{N/A}

- E^{16}: `stni`:{N/A} × `om`:{Case Control, Case Crossover, Case Only, Cohort, Ecologic or Community, Natural History, Other}
- E^{17}: `stni`:{Epidemiological study, Observational study, Other} × `om`:{N/A}

Finally, we have provided some example data in Table 1 that have to be used during some of the upcoming exercises.

Table 1: Example data `purpose` table.

row index	`cst`	`dst`	`stni`	`om`	`cp`	`dp`
1	Observational	Non-interventional	Epidemiological study	Case Crossover	Treatment	Screening
2	Interventional	Interventional	N/A	Cohort	Treatment	Treatment
3	Observational	Non-interventional	Observational study	Natural History	Other	Prognosis
4	Interventional	Non-interventional	Other	N/A	Other	Diagnostic

Exercise 1: Data exploration

Provide solutions to the following data exploration exercises related to the data in the `purpose` table (so not the example data given in Table 1) in your report.

1. Get the total number of rows.
2. Get the total number of rows in which both attributes `cst` and `dst` take value 'Interventional'.
3. Get the total number of rows in which the attributes `cp` and `dp` take equal, non-`NULL` values (so, different values with the same semantical meaning should not be considered equal). Make sure to check this in a case insensitive way.
4. Get the total number of rows containing at least one `NULL`-value.
5. Get the total number of rows containing at least two `NULL`-values.
6. Get, for each attribute, the total number of rows containing a `NULL`-value for the corresponding attribute.

Exercise 2: Basic definitions

Provide answers to the following exercises in your report.

1. Explain in your own words the underlying meaning of edit rules E^3, E^{12} and E^{16}.
2. List, for each of the following edit rules, which attributes are involved in the edit rule.
   a) E^1
   b) E^5
   c) E^{14}
   d) E^{17}
3. For each of the following sets of attributes, list those edit rules that involve all attributes in that set.
   a) {`stni`}
   b) {`cp`,`dp`}
   c) {`dst`,`om`}

Exercise 3: Redundancy

Provide answers to the following exercises in your report.

1. Construct an edit rule E^r that is redundant to E^2 and in which exactly 4 attributes are involved.
2. Is it possible to construct an edit rule E^r that is redundant to E^{14} in which attribute `stni` is not involved? Why (not)?

Exercise 4: Error detection

Provide solutions to the following error detection exercises related to the example data (Table 1) in your report.

1. List, for each row given in Table 1, which edit rules in $\mathcal{E}$ are failed by the corresponding row.
2. Is it possible to give an example row that fails both edit rules E^5 and E^8, but no additional ones. If so, give such an example row. If not, explain why.
3. Is it possible to give an example row that fails both edit rules E^{13} and E^{15}, but no additional ones. If so, give such an example row. If not, explain why.

Provide solutions to the following error detection exercises related to the data in the `purpose` table (so not the example data) in your report. For this, we ask to create a .rbx file containing all the given edit rules defined on the data first. Add this .rbx file to the final .zip file.

4. Get the total number of rows failing at least one edit rule.
5. Get the total number of rows failing edit rule E^{11}.
6. Get the total number of rows failing edit rules E^{1} and E^{17}.
7. List all rows failing the highest number of edit rules? How many edit rules are failed by these rows?
8. List all edit rules that are failed by the highest number of rows? How many rows fail these edit rules?

Exercise 5: Error localization

Provide solutions to the following error localization exercises related to the example data (Table 1) in your report.

1. List, for each row given in Table 1, all the minimal set covers of the edit rules in $\mathcal{E}$ failed by the corresponding row. You may assume that the weight to change an attribute is 1.
2. Indicate, for each minimal set cover of a failing row, whether it is a correct solution to the error localization problem for the corresponding row.

Exercise 6: Implication

Indicate which of the following types of rules you can retrieve by using generator `cp`, a contributing set $\mathcal{E}_c$ of exactly two edit rules and the implication procedure described on slide TODO of the theory lecture presentation, starting from the set of edit rules $\mathcal{E}$. Give, for each indicated type of rule, a contributing set $\mathcal{E}_c$ which leads to a rule of this type after implication by means of generator `cp`. Write your answers in the report.

- tautology
- implied, non-essentially new edit rule
- essentially new edit rule
- contradiction

Exercise 7: Sufficient set generation

During this exercise, we will ask to generate a sufficient set of rules $\underline{\Omega}(\mathcal{E})$ starting from $\mathcal{E}$. As you saw in the theory lecture and the lab session, this can be done by applying the FCF algorithm. However, do not write down the entire execution of FCF in your report as this will require too much space and time, but follow the steps listed below. Assume that the order of the attributes visited in the FCF tree is $1 = \texttt{cp}$, $2 = \texttt{stni}$, $3 = \texttt{dp}$, $4 = \texttt{cst}$, $5 = \texttt{dst}$, $6 = \texttt{om}$.

1. Obviously, start with node (1). List in your report (i) which edit rules are selected as possible contributors and (ii) which essentially new rules are generated from these edit rules in node (1). List for each rule that is generated in either the root node or node (1) whether it is redundant or not. If a rule is redundant, list which edit rules are dominating this rule.

2. Apply the same procedure for node (12) and (123) and write your results in the report.

3. Do you need to visit node (1234) during the remainder of the FCF algorithm? And node (124)? Why (not)?

4. After termination of the FCF algorithm, a sufficient set of edit rules $\underline{\Omega}(\mathcal{E})$ is returned. List all edit rules that are in $\underline{\Omega}(\mathcal{E})$ but not in $\mathcal{E}$ in your report. Add a .rbx file in which all rules in $\underline{\Omega}(\mathcal{E})$ are defined to your final .zip file.

5. List, for each row given in Table 1, which edit rules in $\underline{\Omega}(\mathcal{E}')$ are failed by the corresponding row and all the minimal set covers of the edit rules in $\underline{\Omega}(\mathcal{E}')$ failed by the corresponding row. Because a sufficient set was generated, these minimal set covers are now correct minimal solutions to the error localization problem for the given rows (check this!). You may assume that the weight to change the value of an attribute is 1. Write your answer in the report.

Exercise 8: Imputation

A very straightforward imputation method is called *donor imputation*. Suppose that a row r in a dataset fails some edit rules resulting in a minimal solution S. Donor imputation is going to search for a repair r' (the donor) in the same dataset that (1) does not fail any edit rules, (2) has at least different values for the attributes in S compared to r and (3) has as many equal values for the other attributes compared to r (i.e. the donor row r' should be as closely as possible to the original row r).
Search, for each row given in Table 1, a donor row r' that exists in the `purpose` table of the `clinicaltrials` database. Therefore, consider one solution S for each failing row r (cfr. exercise 7.5) and make sure that r' has at most one different value for the attributes that are not in S compared to r. List, for each failing row, a potential

donor row that meets the above requirements in your report. If no donor is found, mention this too. Write your answers in the report.

Appendix 1: Description of the dataset

In the following, an overview is given of all attributes and the corresponding values that appear in the `purpose` table of the `clinicaltrials` database. Also, notice that we listed the abbreviated version of the attributes between brackets, which we will use during the assignment. For the attributes, we have provided a short explanation of their meaning. Read carefully through this appendix before starting the exercises.

- `ctgov_study_type` (`cst`): Type of study, as reported by ClinicalTrials.gov.
  - Interventional
  - Observational
  - Observational [Patient Registry]
- `drks_study_type` (`dst`): Type of study, as reported by DRKS.
  - Interventional
  - Non-interventional
- `study_type_non_interventional` (`stni`): Type of non-interventional study, as reported by DRKS.
  - Epidemiological study
  - N/A
  - Observational study
  - Other
- `observational_model` (`om`): General design of the strategy for identifying and following up participants during an observational study, as reported by ClinicalTrials.gov.
  - Case Control
  - Case Crossover
  - Case Only
  - Cohort
  - Ecologic or Community
  - N/A
  - Natural History
  - Other
- `ctgov_purpose` (`cp`): Purpose of study, as reported by ClinicalTrials.gov.

  - Basic Science
  - Diagnostic
  - Health Services Research
  - Other
  - Prevention
  - Screening
  - Supportive Care
  - Treatment

- `drks_purpose` (`dp`): Purpose of study, as reported by DRKS.
  - Basic research/physiological study
  - Diagnostic
  - Health care system
  - Other
  - Prevention
  - Prognosis
  - Screening
  - Supportive care
  - Treatment