SEARCH FOR CRITICAL EVENTS IN LEARNING
PART II
ASSOCIATION RULES

June 7, 2019

ABDELMOUMENE Djahid
Université de Cergy-Pontoise

Contents

0.1 Introduction 2

0.2 Objectives 2

0.3 Tools used 2

0.4 Joining the database tables 3

0.4.1 EPM Dataset 3

0.4.2 OULA Dataset 4

0.5 Data imputations 5

0.5.1 EPM Dataset 5

0.5.2 OULA Dataset 5

0.6 Reshaping the data 6

0.6.1 EPM Dataset 6

0.6.2 OULA Dataset 8

0.7 Discretisation 8

0.8 Conclusion 9

0.1 Introduction

This the second report in a set of three that describe the main part to the project, this one explains in detail the process of mining the association rules from the SQL databases.

The two datasets were both extracted from students and their interactions with two different virtual environments.

The first dataset is called Educational Process Mining or EPM for short, mining data from a virtual environment called Deeds (Digital Electronics Education and Design Suite).

The second one called Open University Learning Analytics Dataset (OULAD) containing data extracted from the Virtual Learning Environment (VLE).

0.2 Objectives

The first objective was to transform the datasets into a suitable format, that allows fast manipulation of the content. That format is a relation-based SQL database, this step is necessary because of the size of the data and the complexity of the operation performed on it, more details about this part can be found in the other report.

The main objective was to extract association rules from both of these datasets. These association rules have to be one to one relation found within the data after reshaping and adjusting each database's tables to find the rules.

0.3 Tools used

To mine these association rules various methods and tools were used. **Python3** was used because it has many libraries to manipulate data. Among these libraries *psycopg2* to interact with the *PostgreSQL* database, *Numpy* and *Pandas* to reshape, transform and manipulate the data.

To find the association rules two different methods were used. The first is *Orange* in GUI format, which has widgets to connect to the database and manipulate and reshape the data, and most importantly it has add ons that extend its functionality mainly the one that contains tools to find the association rules.

The second method was in python script format with *mlxtend* which has among many other tools, implementation for calculating association rules and frequent item sets.

0.4 Joining the database tables

Before calculating the association rules the data in their database form needs to get unified into one single table that capture the data of the entirety of the database, the relevant attributes need to be included in this resulting relation. The information loss (the relations between the different tables) ratio needs to be minimal, in order to minimize the filling and imputation for the missing data at the end.

This operation would certainly result in some missing data that even though is minimal, has to be addressed. this process is called imputation and is an unavoidable data quality problem.

0.4.1 EPM Dataset

For this first data set the resulting relation needs to contain the students' grades as well as the data about their interactions with the Deeds virtual environment.

The information for the grades is stored in two different tables, **intergrade** and **finalgrade**. Where the latter contains the results of the last exam in the form of the partial grades for each question in (which were formatted in a more accessible format as explained in the other report). As for the former table, it contains the data of each student grade for each session, that is 6 in total (for any particular students).

As for the virtual environment's data, it is stored in a single table called **Epm** containing various information but mainly the times and dates of actions performed on the Deeds virtual environment, as well as the logs of the mouse and keyboard events, which were abstracted (by the dataset publishers) a little to respect the privacy of the students.

This abstraction was achieved by taking slices of times for the action instead of each individual mouse or keyboard action, that means we only get the total of the clicks, movements or presses in the specified time interval, for example between 10:00:00 and 10:00:50 there were 10 presses and the mouse moved 300 pixels.

This table also contains data about the programs and questions being treated for each separate action, for example the text editor program to solve question number 1 in the first session.

After taking all these criteria into consideration the resulting SQL command looked like this:

```
SELECT
start_time, end_time, epm.student_id, epm.session_id, i_grade,
question_no, max_grade, grade, passage_no,exercice, activity,
mouse_wheel, mouse_wheel_click, mouse_click_left, mouse_click_right,
mouse_movement, keystroke
        FROM
        (SELECT finalgrade.student_id, finalgrade.session_id,
        intergrade.gradeAS i_grade, question_no, max_grade,
        finalgrade.grade, passage_no
                FROM
                intergrade RIGHT JOIN finalgrade
                ON
                intergrade.student_id = finalgrade.student_id AND
                intergrade.session_id = finalgrade.session_id
        ) AS grades
RIGHT JOIN epm
    ON
    grades.student_id = epm.student_id AND
    grades.session_id = epm.session_id;
```

0.4.2 OULA Dataset

For this dataset regardless of its enormous size (500Go uncompressed), the data itself didn't have much detail to it compared to the previous dataset, that is there weren't as many useful attributes that could have been used in the final association rule mining.

Three tables were joined in this process **StudentInfo**, **StudentVle** and **Vle**, the first one because it contains the most detailed information about the students such as the gender, region, grade.. which are valuable attributes when looking for interesting correlations in the data.

The second table **StudentVle** contains the most information (+10M rows) which adds some more credibility and weight to the resulting rules, it also contains the interactions with the virtual environment even though they are very simplified.

The third table contains data that is needed necessarily by the second table to add names of the programs or sites used for each event as well as some important time, date attributes and others that connects it to the courses studied by each student, so it adds more dimensionality to the data.

```
SELECT *
FROM
    (SELECT
    final_result, id_site, sum_click
    FROM
        studentInfo RIGHT JOIN studentVle
        ON
        studentInfo.id_student = studentVle.id_student
    ) AS info_vle
    LEFT JOIN vle
    ON vle.id_site = info_vle.id_site;
```

0.5 Data imputations

The result from the joins of the databases' tables usually results in some missing values if we want to keep the data loss minimal. This is a data quality problem that needs to be addressed before proceeding to the next step.

0.5.1 EPM Dataset

The resulting joined table contained 2% missing (null) values mainly situated in the grades, because some students were not present in the intermediate sessions or didn't pass the final exam in the first or second time.

The grades for the students who were absent in the intermediate session get a 0 for it, but for the final exam there were two passages, so the maximum of the two grades was taken or 0 if the student wasn't present in neither or them.

This process of taking the max between the two was explained in the data description sheet, but the 0 for the missing values was improvised, seeing as the sheet didn't mention anything about this. that is because it seems like this method would be the most logical and that would reflect the student's effort or lack thereof.

0.5.2 OULA Dataset

The combined relations contained 5% missing values situated mainly in the Vle table's part of the join.

The resulting missing values were replace with a blank activity to express the lack of usage of any program during the action, these types of events were not ignores sice they can be used to measure different things but mainly the amount of time each student spends doing actual work (or not), this however is just an observation and was neither exploited nor confirmed.

0.6 Reshaping the data

The resulting table after the imputation of the missing data still contains some either useless or redundant data (As far as the search for the critical events goes). and After taking this into consideration some changes to the attributes were made, and in some occasions new ones were creates from the existing data.

The final attributes should be better suited to the operation at hand, and thus they should give more interesting and readable results.

0.6.1 EPM Dataset

Please note that in the following list the three paragraphs (marked with - each) explain; the significance and meaning of the attribute first, and then the way it was calculated, and also how it can be used when observed in an association rule. In that exact order.

For this dataset the following attributes were the result:

- **Grade**:

  -This attribute is supposed to represent the overall grade of each student, taking into account all of their grades (intermediate and final) into account.

  -After having calculated the final result from by adding the points of each question in the **finalgrade** table, the result is added to the mean after normalizing it (between 0 and 1) of the intermediate grade (based on each session's coefficient). The average of these former two is take and the result is between 0 and 1, where 0 is a complete fail and 1 is the perfect score.

  -This resulting value should directly reflect each student's performance and score, which is useful to have in an association rule because it can give us an idea on the practices each student' category (high, medium and lower grades) in their interaction with the virtual environment.

- **Exercice**:

  -This attribute represents the session and question the students were solving for any particular action while using the virtual environment.

  -Not much was changed from this variable except for some minor adjustments to the format; some values (Es) that are supposed to mean the student wasn't solving any particular question.

  -The value of this attribute can give an idea on the content or rather the effort and tools (programs) needed to solve any particular question, all of these information could be deduced from an association rule containing this attribute.

- **Activity**:

  -This one serves to show the active program for any particular set of actions (mouse, keyboard).

  -The changes to the value were insignificant, other than the removal of some redundant information at the end of some values (values that end in the question and session number, Deeds_Es_1_2 would be changed to Deeds for example).

  -The appearance of this attribute in an association rule would suggest that the user used X program to do Y action, we can deduce from such information that certain programs can help in some other action more than other (programs).

- **Mouse_wheel**:

  -This one signifies the movement of the mouse wheel for any particular event.

  -No changes except for the discretization (which will be discussed later) were made.

  -The value of this attribute could signify the need for the usage of the mouse or (mouse wheel in particular) for any action.

- **Mouse_wheel_click**:

  -This one signifies the number of click for the middle mouse button (wheel) for any particular event.

  -No changes except for the discretization (which will be discussed later) were made.

  -The value of this attribute could signify the need for the usage of the mouse or (mouse wheel in particular) for any action.

- **Mouse_click_left**, **Mouse_click_right, mouse_movement**:

  -These three signify the actions of the mouse buttons and movement for any particular event.

  -No changes except for the discretisation (which will be discussed later) were made.

  -The value of these attribute could signify the need for the usage of the mouse for any action.

- **Keystroke**:

  -This attribute signify the number of the keyboard presses for any particular event.

  -No changes except for the discretisation (which will be discussed later) were made.

  -The value of these attribute could signify the need for the usage of the keyboard for any action.

- **Duration**:

  -This attribute signifies the number of seconds taken by any particular event.

  -This variable was calculated using the difference between **start_time** and **end_time**, which gives the total amount of seconds elapsed in the action.

  -This could help show the sort of actions that take more (or less) time to achieve.

0.6.2 OULA Dataset

The work on transforming this dataset has yet to conclude, although a lot of the methods and tools that were used on the EPM datasets will be reused for this one, especially the discretizing of the numerical values.

0.7 Discretisation

Certain numerical values had to be discretized in order to pass them to the appropriate tools, to find the final association rules and frequent itemsets. There are many ways this can be achieved.

One of the methods that can be used to achieve this, is the equal length interval discretisation, meaning that intervals of equal length (between the min and max values) are generated, and then each value is transformed into the appropriate bin (interval).

There are many problems with this method that can affect the quality of the result, the most important of which is that the bins can be very unbalanced, meaning some of them could contain the most of the values whereas the other bins contain very few of them, this is caused because most of the attributes used (mouse clicks for example) follow a normal distribution, which means most values approach a certain mean value, and very few of them are far away from the mean to be classified under some of the more extreme intervals.

A second method would be the frequency-based discretization, which means that we choose intervals that contain more or less the same amount of values, meaning that if we discretize into 4 bins, each of these bins will contain 25% of the value.

This method better suits the data at hand because the attributes are normally distributed.

0.8 Conclusion

The search for the association rules will later be import when looking for critical events in the datasets, the results of the data cleaning and discretizing was saved into two CSV files one in normal format (values are strings) and the second in a dummified version (meaning each different value is transformed into an attribute in the resulting csv) and the values become either 1s or 0s signifying the presence or absence of the attribute.

These files can be directly used to generate the association rules and the frequent itemsets, with customizable minimal support and confidence (along with other criteria).