**A report on**

**INFO-H423 - Data Mining. Assignment – part 2.**

By

Raymond Lochner (000443637)

Aleksei Karetnikov (000455065)

Aldar Saranov (000435170)

# Dataset preprocessing

Data is loaded from the Excel file. When reading the data, the **ticket** attribute must be marked as ID.

While performing a brief input data analysis, one can remark that no missing values exist in the dataset, therefore no row cleansing is required.

Before applying FPGrowth itemset computation, one has to convert all values in the dataset into binomials (operator requirement).

**ITOwner**, **Requestor**, **Days** and **ticket** are the only integer attributes. However, it only makes sense to discretize the **Days** attribute, since if we discretize the **ITOwner**, **Requestor** and **ticket** (various IDs), then the obtained ranges most likely will not be implicated with the satisfaction. Eventually it was decided to discretize ticket into 5 bins. Too many discretization ranges may lead to overspecialization of the obtained rules.

**ticket** will not contribute the items at all, since it is just a unique request identifier and cannot affect the construction of association rules. **ITOwner** and **Requestor** however may participate the rule construction. One problem is how to represent them, since converting them to binaries will generate 2000 **Requestor** and 50 **ITOwner** binary flags. Naturally, it may happen that for example a certain requestor irrationally estimates the work of the tech support. Or on contrary, a certain tech support employee does not handle his requests very well, and thus causes negative satisfiability. Eventually it was decided to remove the **Requestor** due to very large number of its values, but to leave the **ITOwner**.

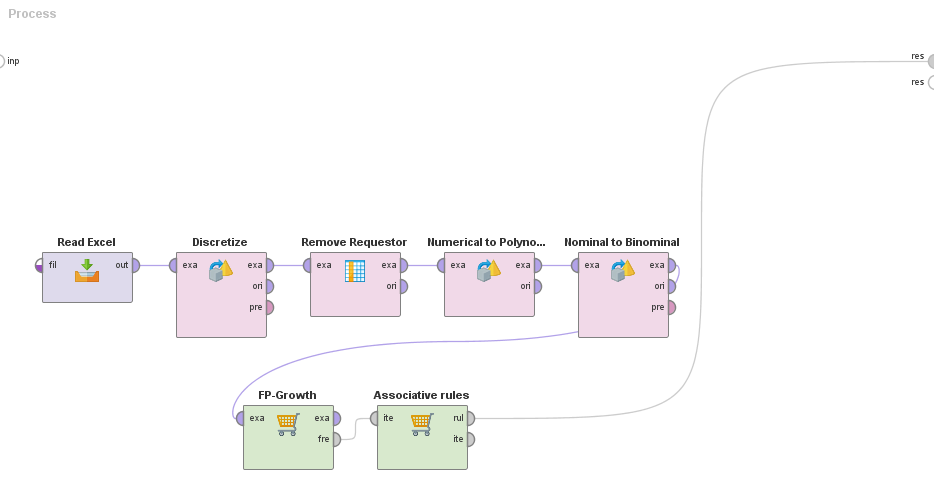
**ITOwner** is mapped into polynomial first by applying “Numerical to Polynomial”, and then decomposed into binary flags by applying “Nominal to Binomial”.

# Data mining

Before generating rules one has to obtain frequent itemsets. This is done by means of “FP-Growth” operator which aimed to calculate all of them by using FP-tree. According to the operator’s features all the input attributes must be binominal.

We set the **max items** equal to 3 because we don’t need rules which have more than 3 itemsets and **min** support = 0.01. We have selected such value because our dataset contains 100.000 records and 50. So, the average percentage of requests per each IT specialist in not significantly more than 2% of all the tickets. Then, we can assume that minimal support should be no less than half of the tickets s=0.02/2=0.01.

The next step is generating of the association rules with the corresponding operator by confidence. After the series of experiments which are described below in the section Model evaluation, we have concluded that the appropriate **min confidence** c=0.3.



*Figure 1. Data mining flow.*

# Model evaluation

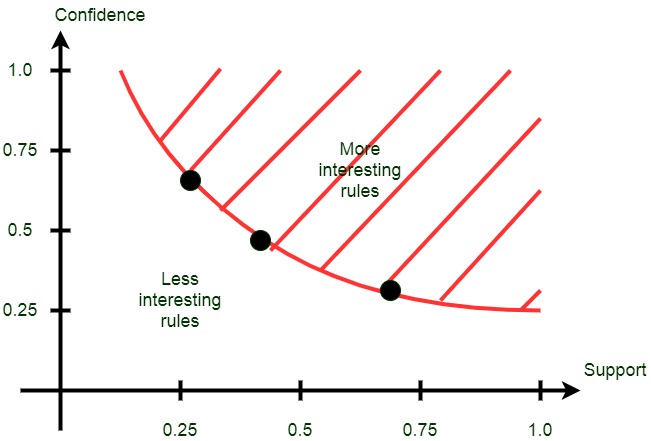
The process completes in 3s.

We need to obtain the most interesting rules related to the attributes **Satisfaction** and **daysOpen**. One can remark following criterias of interest:

1. Support.
2. Confidence.
3. Generality of the rule (low number of rule items).

Min support and generality (max items) are adjusted in “FP-growth” operator. Min confidence is adjusted in “Associative rules” operators. Hereby, we have a multivariate optimization problem.

Another feature of this problem is that by making the constraint more strict one can obtain less rules related to a particular item (e.g. **Satisaction** items or **daysOpen** items). Conversely, if one adjusts the constraints to less strict, one can obtain a bunch of rules, many of those tend to be too specific. Therefore, the configurations should be adjusted depending on what kind of result do we obtain.



*Figure 2. Optimization problem illustration.*

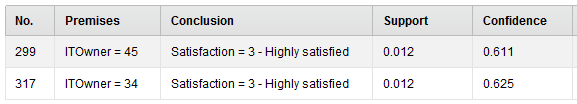
One also can remark that by forcing support larger than 0.02 will likely to cut-off all the ITOwner related rules, since in average only 0.02 of all requests are assigned to a particular employee. In addition some lower bound of confidence should be ensured and max number of rule items. We have chosen max item = 3, since rules with item number >= 4 are difficult to analyze from the human point of view.

In order to find the configuration that is located at subjective margin of the interesting area we have conducted series of rule mining experiments, with varying support and confidence parameters located along some configuration set.

The proposed configuration set was following:

|  |  |  |  |
| --- | --- | --- | --- |
| Number | Support | Confidence | Max item |
| 1 | 0.012 | 0.4 | 3 |
| 2 | 0.010 | 0.3 | 3 |
| 3 | 0.008 | 0.2 | 3 |

*Table 3. Highly-satisfied related rules in configuration-1.*



*Figure 4. Highly-satisfied related rules in configuration-1.*

We won’t choose configuration-1 because it provides too few and very general rules.

**

*Figure 5. Highly-satisfied related rules in configuration-2.*

Configuration-2 provides quite interesting rules, balanced between its support, confidence and generality.

**

*Figure 6. Highly-satisfied related rules in configuration-3.*

Configuration-3 provides a large number of weak-performing rules, that are among other things too specific.

Finally, we can state that the configuration-2 is more suitable for analysis.

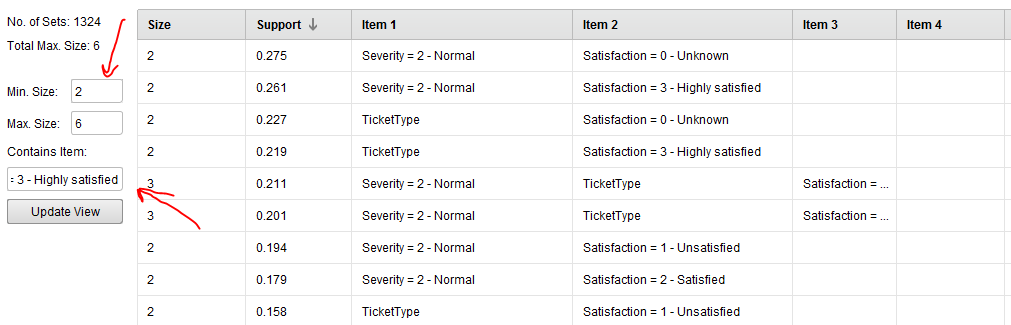
# Analysis

Browsing the rules

In order to view the most interesting itemsets that were obtained by applying FP-Growth one can set filters:

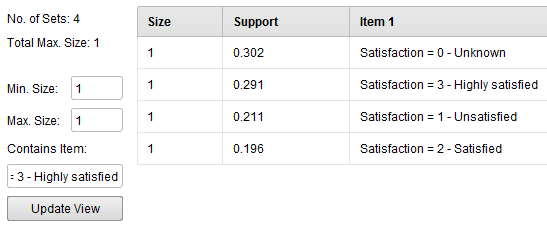
* Min. Size = 2,
* Contains Item = “Satisfaction = 0 - Unknown, Satisfaction = 1 - Unsatisfied, Satisfaction = 2 - Satisfied, Satisfaction = 3 - Highly satisfied”

This is done because we only want to obtain information related to the satisfiability of the employees and that does not contain only one item.



*Figure 7. Rule filtering*

If we examine one-itemsets then we will only get the distribution of the satisfiability over its possible values.



*Figure 8. 1-itemsets distribution.*

Satisfaction conclusions

For satisfactory=3, some conclusions can be made by Figure 5.

We disregard several redundant rules like:

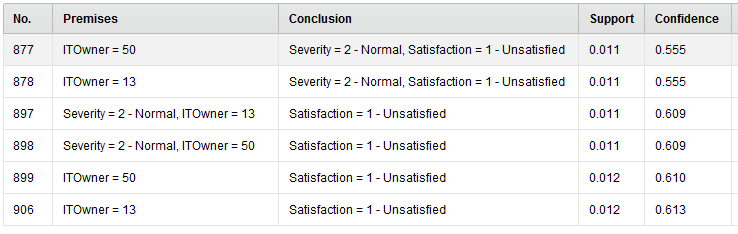
* **<TicketType and Severity=3> → <Satisfaction=3>** is not used, **<Severity=3> → <Satisfaction=3>** is used instead.
* **<Severity=2 and ITOwner=21> → <Satisfaction=3>** is not used, **<ITOwner=21> → <Satisfaction=3>** is used instead.
* **<Severity=2 and ITOwner=45> → <Satisfaction=3>** is not used, **<ITOwner=45> → <Satisfaction=3>** is used instead.
* **<Severity=2 and ITOwner=34> → <Satisfaction=3>** is not used, **<ITOwner=34> → <Satisfaction=3>** is used instead.

From the rest of the rules we can clearly state that the employees #34. #45, #21, #27 are working rather well, so their level of satisfaction is high.

The satisfactory level of major problems is higher than the satisfactory level of other problem complexities.

For satisfactory=2, we have not managed to obtain any rules due to our constraints of support and confidence.

For satisfactory=1, we have rules according to Figure 9.



*Figure 9. Unsatisfied related rules in configuration-3.*

Similarly we may disregard the redundant rules and tell that employees #13 and #50 are low-performing.

Finally, we have received a set of rules for attributes which correspond to length of tickets and user’s satisfaction, which we need to find according to the task.

1. Quickness of the support.

At the step of data preprocessing we have grouped our dataset’s attribute **daysOpen** and received 5 groups: the problems which are resolved in the same day, 1-3, 4-6, 7-13 and more than 14 days. We have received the following sets of the rules which we need to recognize to get the most interesting and appropriate associations (we have removed the records with Satisfaction in Itemset because we are recognizing Satisfaction as a resulting metrics).

The first day:

|  |  |  |
| --- | --- | --- |
| **Itemset** | **s** | **c** |
| FiledAgainst = Access/Login, Priority = 1 - Low | 0,035 | 0,690 |
| FiledAgainst = Access/Login, RequestorSeniority = 1 - Junior | 0,043 | 0,724 |
| FiledAgainst = Access/Login, Severity = 3 - Major | 0,011 | 0,741 |
| FiledAgainst = Access/Login, Priority = 2 - Medium | 0,037 | 0,765 |
| RequestorSeniority = 2 - Regular, FiledAgainst = Access/Login | 0,095 | 0,774 |
| FiledAgainst = Access/Login | 0,236 | 0,787 |
| Priority = 0 - Unassigned, FiledAgainst = Access/Login | 0,072 | 0,788 |
| TicketType, FiledAgainst = Access/Login | 0,177 | 0,788 |
| Severity = 2 - Normal, FiledAgainst = Access/Login | 0,214 | 0,789 |
| FiledAgainst = Access/Login, RequestorSeniority = 3 - Senior | 0,047 | 0,824 |
| Priority = 3 - High, FiledAgainst = Access/Login | 0,092 | 0,842 |
| FiledAgainst = Access/Login, RequestorSeniority = 4 - Management | 0,050 | 0,844 |

*Table 9.*

According to the itemsets, the main type of the problem which could be resolved at the same day is Access/Login problems. Different support of rules with different requestor seniority corresponds to the number of employees of such grades.

1-3 days

|  |  |  |
| --- | --- | --- |
| **Itemset** | **s** | **c** |
| FiledAgainst = Access/Login, Priority = 1 - Low | 0,016 | 0,309 |
| FiledAgainst = Software | 0,064 | 0,319 |
| FiledAgainst = Software, RequestorSeniority = 3 - Senior | 0,013 | 0,348 |
| FiledAgainst = Software, RequestorSeniority = 4 - Management | 0,014 | 0,375 |
| Priority = 0 - Unassigned, FiledAgainst = Software | 0,019 | 0,317 |
| Priority = 3 - High, FiledAgainst = Software | 0,028 | 0,382 |
| RequestorSeniority = 2 - Regular, FiledAgainst = Software | 0,026 | 0,310 |
| Severity = 2 - Normal, FiledAgainst = Software | 0,058 | 0,317 |

*Table 10.*

In this group, the main itemset is software problem. Some additional parameters, such as different priority, severity and requestor seniority follow to redundancy and so they could be generalized. The second separate ticket type is problem with Access/Login with low priority.

4-6

|  |  |  |
| --- | --- | --- |
| **Itemset** | **s** | **c** |
| FiledAgainst = Systems, Priority = 0 - Unassigned | 0,036 | 0,302 |
| TicketType, FiledAgainst = Systems | 0,091 | 0,305 |
| TicketType, FiledAgainst = Software | 0,048 | 0,321 |
| Priority = 3 - High, FiledAgainst = Software | 0,024 | 0,327 |
| FiledAgainst = Software, RequestorSeniority = 4 - Management | 0,013 | 0,330 |
| FiledAgainst = Systems, RequestorSeniority = 3 - Senior | 0,024 | 0,333 |
| FiledAgainst = Software, RequestorSeniority = 3 - Senior | 0,013 | 0,334 |
| FiledAgainst = Systems, RequestorSeniority = 4 - Management | 0,028 | 0,352 |
| FiledAgainst = Systems, Priority = 3 - High | 0,052 | 0,353 |

*Table 11.*

The vast majority of tickets which can be resolved in 4-6 days correspond to problems with System and Software but with higher priority for Senior and Management requestor seniority. This part looks a little bit more important because it touch problems with higher employee grades and it can follow to lower satisfaction.

7-13

|  |  |  |
| --- | --- | --- |
| **Itemset** | **s** | **c** |
| FiledAgainst = Systems, Priority = 0 - Unassigned | 0,036 | 0,301 |
| RequestorSeniority = 2 - Regular, FiledAgainst = Systems | 0,051 | 0,304 |
| FiledAgainst = Systems | 0,122 | 0,305 |
| Severity = 2 - Normal, FiledAgainst = Systems | 0,112 | 0,309 |
| FiledAgainst = Systems, Priority = 2 - Medium | 0,020 | 0,311 |
| FiledAgainst = Systems, RequestorSeniority = 3 - Senior | 0,023 | 0,312 |
| RequestorSeniority = 2 - Regular, FiledAgainst = Software | 0,026 | 0,312 |
| FiledAgainst = Systems, RequestorSeniority = 4 - Management | 0,025 | 0,313 |
| FiledAgainst = Systems, Priority = 3 - High | 0,047 | 0,317 |
| FiledAgainst = Software, Priority = 2 - Medium | 0,011 | 0,318 |
| TicketType, FiledAgainst = Software | 0.051 | 0,340 |
| TicketType, FiledAgainst = Systems | 0,103 | 0,343 |
| FiledAgainst = Software, RequestorSeniority = 1 - Junior | 0,015 | 0,377 |
| RequestorSeniority = 2 - Regular, FiledAgainst = Hardware | 0,017 | 0,422 |
| FiledAgainst = Software, Priority = 1 - Low | 0,014 | 0,430 |
| FiledAgainst = Hardware | 0,044 | 0,445 |
| Priority = 0 - Unassigned, FiledAgainst = Hardware | 0,013 | 0,448 |
| Severity = 2 - Normal, FiledAgainst = Hardware | 0,041 | 0,452 |
| TicketType, FiledAgainst = Hardware | 0,036 | 0,481 |
| RequestorSeniority = 4 - Management, FiledAgainst = Hardware | 0,011 | 0,547 |
| Priority = 3 - High, FiledAgainst = Hardware | 0,020 | 0,555 |

*Table 12.*

The majority of ticket which could be resolved in 7-13 days has problems with Hardware and System. It also can be generalized.

14+ days

|  |  |  |
| --- | --- | --- |
| **Itemset** | **s** | **c** |
| TicketType, FiledAgainst = Systems | 0,106 | 0,352 |
| FiledAgainst = Hardware | 0,048 | 0,479 |
| Severity = 2 - Normal, FiledAgainst = Hardware | 0,043 | 0,475 |
| TicketType, Priority = 1 - Low | 0,042 | 0,331 |
| TicketType, FiledAgainst = Hardware | 0,039 | 0,519 |
| FiledAgainst = Systems, RequestorSeniority = 1 - Junior | 0,030 | 0,376 |
| FiledAgainst = Systems, Priority = 1 - Low | 0,030 | 0,427 |
| RequestorSeniority = 2 - Regular, FiledAgainst = Hardware | 0,021 | 0,511 |
| FiledAgainst = Systems, Priority = 2 - Medium | 0,020 | 0,310 |
| Priority = 0 - Unassigned, FiledAgainst = Hardware | 0,014 | 0,472 |
| Priority = 3 - High, FiledAgainst = Hardware | 0,013 | 0,350 |
| RequestorSeniority = 1 - Junior, FiledAgainst = Hardware | 0,012 | 0,618 |
| Priority = 1 - Low, FiledAgainst = Hardware | 0,012 | 0,684 |

*Table 13.*

The longest tickets correspond to different hardware and system problems with the lowest priority for junior grades.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | 1-3 | 4-6 | 7-13 | 14+ |
| {FiledAgainst = Access/Login} | {FiledAgainst = Software}  {FiledAgainst = Access/Login, Priority = 1 – Low} | {FiledAgainst = Systems RequestorSeniority = 4 - Management (OR 3 – Senior)}  {FiledAgainst = Software RequestorSeniority = 4 - Management (OR 3 – Senior)} | {FiledAgainst = Systems}  {FiledAgainst = Hardware} | {FiledAgainst = Hardware} |

*Table 14.*

Consequently, we can see that the main diversification reason is type of the problem. Sometimes it includes some extraordinary values, like in group 4-6 days, which contents some problems with software and system.

General conclusion

We failed to find a direct dependency between time of the problem handling and the satisfaction level.

However, we have found many relations between satisfaction levels and in certain employees. So, certain measures can be undertaken in order to increase the quality of IT support (e.g. such as activity changing, work optimizing, financial sanctions or even firing).