FRAGILE STATE INDEX ANALYSIS

Chirag Jain 801026633

Hozefa Haveliwala 800936900

Narahari Battala 801018499

Shubhra Mishra 801024865

ITCS -6162: Knowledge Discovery in Database

TABLE OF CONTENTS

INTRODUCTION

Fragile State Index is an annual report prepared to gauge the vulnerability to conflict or collapse. It ranks the countries that are the member of the UN. This index is compiled using 12 indicators each on a scale of 0-10. The total from 0-120, then helps in ranking the countries. Scores are obtained via a process involving content analysis, quantitative data, and qualitative review. All the countries are then classified into one of the four categories i.e. Alert, Warning, Stable and Sustainable.

AIM OF THE STUDY

The aim of the study is to apply different classifier to the dataset obtained from the Fragile State Index data and compare their precision in correctly classifying the countries. Additionally, to increase the precision of classification 6 additional indicator are needed to be collected and merged into the existing data. This addition helps in improving classifier models' accuracy. Lastly action rules are to be extracted from the data to help identify key indicators for the countries the are in the Alert category that will assist in transitioning these countries into a much safer category. Software packages like WEKA, LISP Miner etc can be used to classifier/discretize/mine the datasets.

STEPS PERFORMED FOR THIS STUDY

The data that is gathered, consists of the years from 2010 through 2013. It consists of 12 world development indicators which determined the initial Fragile States Index of the countries. The total number of countries are around 177. Those indicators affect the Fragile Index count of the country for giving the rank of the country in the world.

The next step is to add 6 more indicators which would help increase the accuracy of the resulting Fragile States Index. Now the 6 indicators were appended to the previous data of 12 indicators for all the countries in the given period of 2010-2013. The data was added to the 12 indicators and a total was calculated for each country in each year. The data was discretized in the software DataPreparator which categorized it in four ranges and each range was named according to the present state of the country as 'Alert', 'Warning', 'Stable' and 'Sustainable'. These categories are in the order of the worst country state to a good state.

The discretized and ranged up data was converted to an excel sheet. Then the old and new data was run through different classifiers using the WEKA Tool. Then, the accuracy of the classification models was compared between the old and new data sets.

NEW ATTRIBUTES

Fuel Imports – Every country has its fuel needs which are affected by the level of development. The country has its own sources of the fuel but the requirements are not met every time by its own. Here the import of fuel comes in the scenario and this indicates the amount of fuel consumption. As the need are not met in the country, then the fuel is imported. More imports means more development but also increased expense of the country for obtaining the fuel. This factor determines the country's state of collapse.

Fossil Fuel Energy Consumption (% of total): The dependency on fossil fuels has gained enormous attention in the recent decade as source of pollution. Thus, a country that depends more on coal and natural gas might face challenges related to External Intervention (such as international treaties regarding climate change).

Taxes on International trade (% of Revenue): For a country, the revenue comes from various sources and taxes are one of the major contributors to the country's revenue. The tax obtained from international trade is higher than the domestic trade and determines how good is the status of the country in the world market. For a developed, stable country, the level of international trade is very high which in-turn provides with higher levels of taxes for the country. Now tax collection determines the economic condition of the country and that's why it is taken as one of the indicators for the Fragile Index.

Life Expectancy at Birth (total years): The Life Expectancy at birth, as we see it, is a direct correlation of living conditions in a country. Given poor conditions, deaths are bound to be high, and thus can play a significant role in the countries alertness level. Our contention is that high death rates are closely related to the Public Services indicator, in the capacity that public health services are of high quality and regularly available, and will greatly affect the level of alertness of a country.

Health Expenditure Total (%GDP): To provide good health service and better living conditions, the country's alertness needs to be very high because it is necessary to keep a check on the reason of deaths and the number of deaths per unit person in an area. This is the direct result of the Health Expenditure which the government uses for the improvement of the country's health scenario. IT is a major indicator for alertness of a country.

Global Peace Index: The GPI measure the level of peacefulness among the citizens. It is ought to be high if the citizens are satisfied with the living and working conditions of the society and country in which they are living. The GPI comes first if measuring the peacefulness of a country. A low level of GPI indicates high alertness for the country and high GPI indicate the country is stable and prospering**.**

GATHERING NEW DATA

The data which we have used was gathered from the World Bank website which had the database for all the countries classified in the basis of World Development Indicators. The Fragile States Index data was downloaded from the website of the organization - Fund For Peace. It is a non-profit organization for works to prevent conflict and promote sustainable security globally by building relationships and trust across diverse sectors. The data was chosen in a way that it doesn't has a lot of null, undetermined values so that the result of data processing as accurate and near to the actual value. The indicators were also filtered and only those indicators were used which gives an appropriate image of the country's stability on account of social, political and economic factors. The data after being gathered was scaled so that each set of data can match each other and comparisons can be performed. A total of twelve world development indicators were used initially and finally it was increased to six more i.e. eighteen to improve the accuracy of comparison.

CLASSIFIER INFORMATION

1. **JRip**
   JRip (RIPPER) is one of the basic and most popular algorithms. Classes are examined in increasing size and an initial set of rules for the class is generated using incremental reduced error JRip proceeds by treating all the examples of a particular judgment in the training data as a class, and finding a set of rules that cover all the members of that class. Thereafter it proceeds to the next class and does the same, repeating this until all classes have been covered.

2. **PART**
   It combines the divide and conquer technique with separate and conquer strategy of rule learning. It builds a partial decision tree on the current set of instances. It then creates a decision tree where the leaf is the rule with the largest coverage.

3. **J48**
   This algorithm works by trying to split based on features that give the purest end node, meaning it aims to split such that each feature leads to a uniform decision node with fewest splits. This is a top-down approach that uses information gain at each node to reach the pure nodes (the decision nodes) and form the smallest tree.

4. **Random Forest**
   Random Forest starts out by evaluating how many features our data set has. In our case, we have 11 features. The classifier then generates several decision trees using randomly chosen columns each time to classify the data. The aggregate of the outcome of these trees is what is used to ultimately make a classification decision.
   When building a tree, the algorithm may choose to select a random two features, and out of

those features, splitting until we reach a leaf node with a decision. The result is many trees that are all formed by random 4 set features. The classifier builds these set of random trees (now a random forest) by the supplied training data, which in our case chose 66% of our data. The test data is then used to perform classification. The training data is fed to a random set of trees, and the result classification is noted. The final classification for that instance of data is calculated simply by majority votes.

DISCRETIZATION USING BASE ATTRIBUTES

Ranges Summary for all the years

Year	Alert	Warning	Stable	Sustainable
2010	<90.4-114.3>	<66.5-90.4)	<42.6-66.5)	<18.7-42.6)
2011	<89.98-113.4>	<66.55-89.98)	<43.12-66.55)	<19.7-43.12)
2012	<91.18-114.9>	<67.45-91.18)	<43.73-67.45)	<20_43.73)
2013	<89.925-113.9>	<65.95-89.925)	<41.975-65.95)	<18-41.975)

DISCRETIZATION USING NEW ATTRIBUTES

Ranges Summary for all the years

Year	Alert	Warning	Stable	Sustainable
2010	<109.252_133.036>	<85.469_109.252)	<61.685_85.469)	<37.902_61.685)
2011	<107.361_130.22>	<84.501_107.361)	<61.642_84.501)	<38.783_61.642)
2012	<108.05_131.18>	<84.92_108.05)	<61.79_84.92)	<38.66_61.79)
2013	<109.28_133.4>	<85.15_109.28)	<61.02_85.15)	<36.9_61.02)

CLASSIFIER USING NEW ATTRIBUTES

SUMMARY

Year	Classifier	Accuracy (%)
2010	JRip	83.05
2010	PART	80.79
2010	J48	79.66
2010	Random Forest	83.61
2011	JRip	77.40
2011	PART	79.09
2011	J48	77.40
2011	Random Forest	82.48
2012	JRip	77.5
2012	PART	80.89
2012	J48	80.89

2012	Random Forest	83.70
2013	JRip	76.40
2013	PART	79.77
2013	J48	79.77
2013	Random Forest	83.70

EXAMPLE

The following is an example for the year 2010. The test was done using cross validation using 10 folds.

1. **JRip**

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	147	83.0508 %
Incorrectly Classified Instances	30	16.9492 %
Kappa statistic	0.745	
Mean absolute error	0.1031	
Root mean squared error	0.2824	
Relative absolute error	30.4883 %	
Root relative squared error	68.755 %	
Total Number of Instances	177	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.806	0.055	0.758	0.806	0.781	0.734	0.909	0.672	Alert
	0.872	0.187	0.815	0.872	0.843	0.686	0.846	0.777	Warning
	0.657	0.028	0.852	0.657	0.742	0.697	0.820	0.650	Stable
	0.960	0.007	0.960	0.960	0.960	0.953	0.994	0.915	Sustainable
Weighted Avg.	0.831	0.107	0.833	0.831	0.829	0.734	0.873	0.753	

=== Confusion Matrix ===

```
 a  b  c  d   <-- classified as
25  6  0  0 |  a = Alert
 8 75  3  0 |  b = Warning
 0 11 23  1 |  c = Stable
 0  0  1 24 |  d = Sustainable
```

2. PART

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	143	80.791 %
Incorrectly Classified Instances	34	19.209 %
Kappa statistic	0.7173	
Mean absolute error	0.1004	
Root mean squared error	0.3074	
Relative absolute error	29.6901 %	
Root relative squared error	74.8354 %	
Total Number of Instances	177	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.806	0.048	0.781	0.806	0.794	0.749	0.876	0.688	Alert
	0.826	0.132	0.855	0.826	0.840	0.695	0.839	0.768	Warning
	0.714	0.077	0.694	0.714	0.704	0.630	0.813	0.568	Stable
	0.880	0.026	0.846	0.880	0.863	0.840	0.962	0.793	Sustainable
Weighted Avg.	0.808	0.092	0.809	0.808	0.808	0.712	0.858	0.718	

=== Confusion Matrix ===

```
 a  b  c  d   <-- classified as
25  6  0  0 |  a = Alert
 7 71  8  0 |  b = Warning
 0  6 25  4 |  c = Stable
 0  0  3 22 |  d = Sustainable
```

---

3. J48

=== Stratified cross-validation ===
=== Summary ===

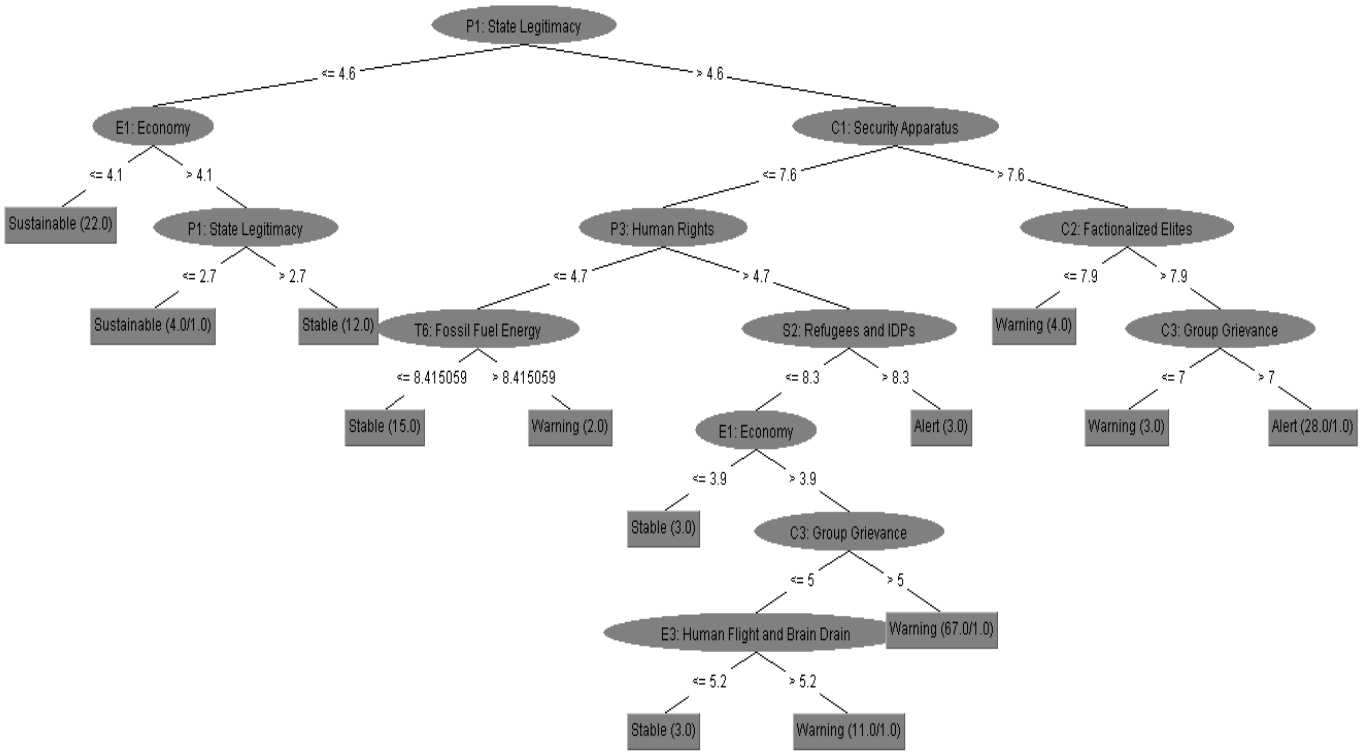
Correctly Classified Instances	141	79.661 %
Incorrectly Classified Instances	36	20.339 %
Kappa statistic	0.7036	
Mean absolute error	0.1081	
Root mean squared error	0.3123	
Relative absolute error	31.967 %	
Root relative squared error	76.0342 %	
Total Number of Instances	177	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.839	0.089	0.667	0.839	0.743	0.687	0.875	0.604	Alert
	0.791	0.121	0.861	0.791	0.824	0.673	0.831	0.757	Warning
	0.743	0.063	0.743	0.743	0.743	0.679	0.836	0.648	Stable
	0.840	0.020	0.875	0.840	0.857	0.834	0.950	0.831	Sustainable
Weighted Avg.	0.797	0.090	0.805	0.797	0.799	0.700	0.857	0.719	

=== Confusion Matrix ===

```
 a  b  c  d   <-- classified as
26  5  0  0 |  a = Alert
13 68  5  0 |  b = Warning
 0  6 26  3 |  c = Stable
 0  0  4 21 |  d = Sustainable
```



Decision tree for J48

4. **Random Forest**

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	148	83.6158 %
Incorrectly Classified Instances	29	16.3842 %
Kappa statistic	0.7438	
Mean absolute error	0.2095	
Root mean squared error	0.2763	
Relative absolute error	61.9595 %	
Root relative squared error	67.2684 %	
Total Number of Instances	177	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.677	0.007	0.955	0.677	0.792	0.773	0.984	0.945	Alert
	0.977	0.242	0.792	0.977	0.875	0.749	0.945	0.894	Warning
	0.629	0.035	0.815	0.629	0.710	0.657	0.956	0.884	Stable
	0.840	0.007	0.955	0.840	0.894	0.880	0.998	0.991	Sustainable
Weighted Avg.	0.836	0.127	0.848	0.836	0.830	0.754	0.962	0.915	

=== Confusion Matrix ===

```
 a  b  c  d   <-- classified as
21 10  0  0 |  a = Alert
 1 84  1  0 |  b = Warning
 0 12 22  1 |  c = Stable
 0  0  4 21 |  d = Sustainable
```

CONCLUSION

With the additions of the new attributes, Random Forest gave the best classification results.

ACCURACY BASED ON NEW CLASSIFIERS

Year	Best Classifier	Accuracy (%)
2010	Random Forest	83.61
2011	Random Forest	82.48
2012	Random Forest	83.70
2013	Random Forest	83.70

NEW ATTRIBUTE ANALYSIS

The new attributes were analysed using the attribute selection option from WEKA. The aim of this was to check which attribute has more effect on the overall accuracy on the classification of the countries. Following is the output from WEKA. We can see that Global Peace Index in the most important attribute while Fossil Fuel Energy is the lowest.

Weka Analysis

=== Run information ===

Evaluator: weka.attributeSelection.InfoGainAttributeEval
Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation: c_fsi-2010-attributerank
Instances: 177
Attributes: 7
New_TOTAL
T1: Global Peace Index
T2: Health
T3: Life Expectency
T4: InternationalTaxes
T5: FuelImports
T6: Fossil Fuel Energy
Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:
Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 1 New_TOTAL):
Information Gain Ranking Filter

Ranked attributes:
0.772 2 T1: Global Peace Index
0.517 4 T3: Life Expectency
0.215 3 T2: Health
0.142 5 T4: InternationalTaxes
0 6 T5: FuelImports
0 7 T6: Fossil Fuel Energy

ACTION RULES USING LISP MINER

After getting the discretized data sets and classifying them, we want to derive action rules based on the values for each of the indicators.

Steps :

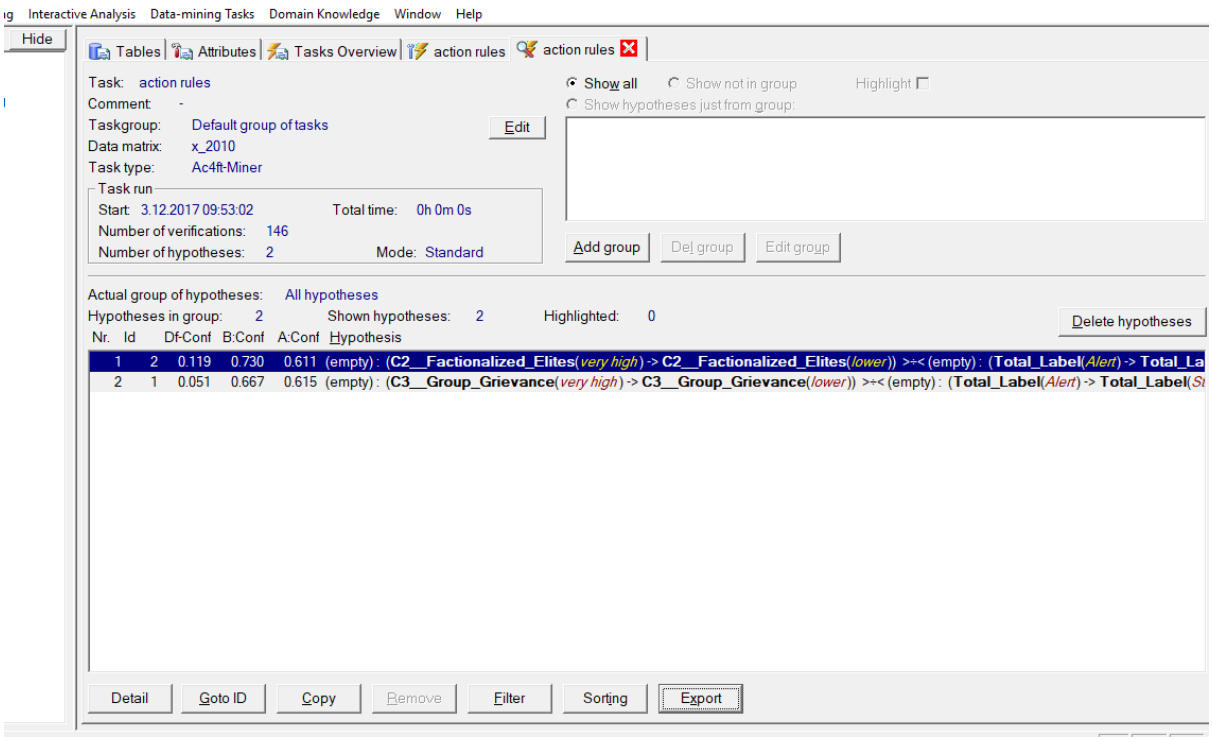
- We merge the old data sets with new attributes and discretize them. Further, we derive the action rules based on these new identifiers.

-We have taken 'Country(name)' as stable attribute, 'Total_Label' as decision attribute and all the other indicators as flexible attributes.

- The action rules generated give a recommendation that what steps/reforms could be opted by a country in order to move to the less fragile state.

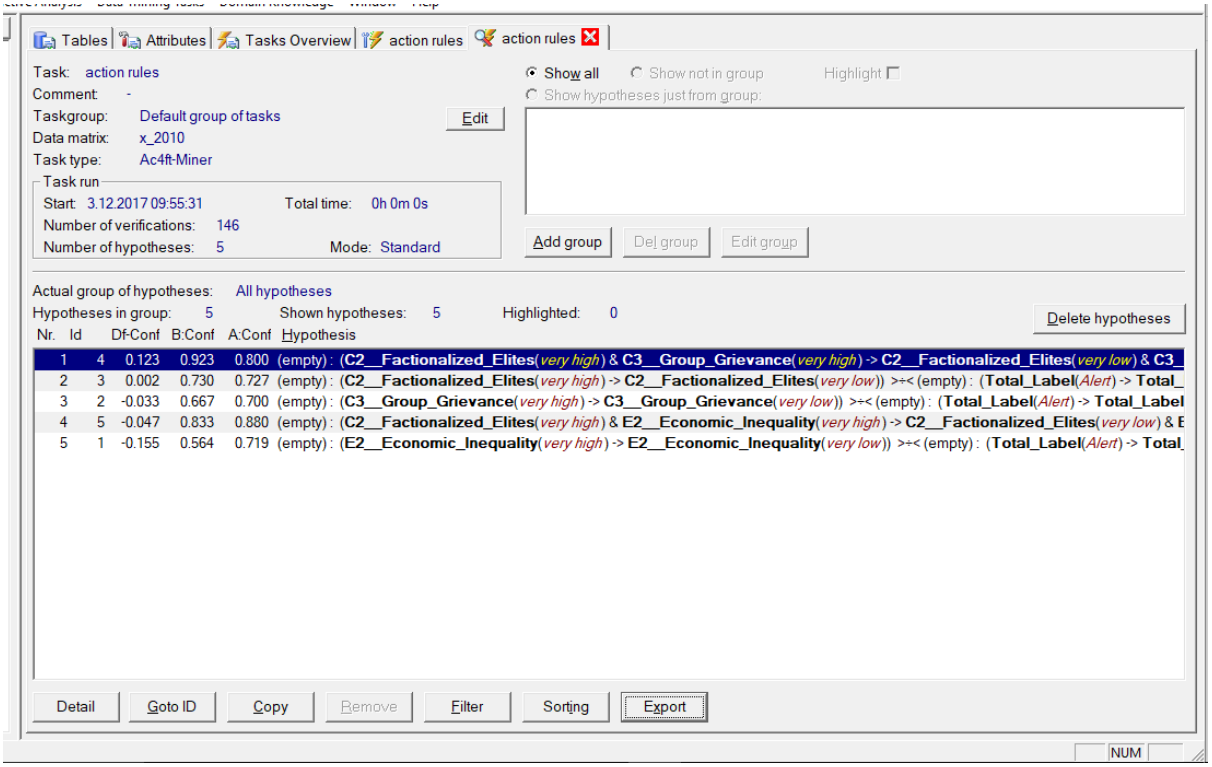
<u>For Year 2010</u>:

- **Action Rules for Alert to Stable:**
  1. (C2__Factionalized_Elites(very high) -> C2__Factionalized_Elites(lower)) >÷< (Total_Label(Alert) -> Total_Label(Stable))
  2. (C3__Group_Grievance(very high) -> C3__Group_Grievance(lower)) >÷< (Total_Label(Alert) -> Total_Label(Stable))

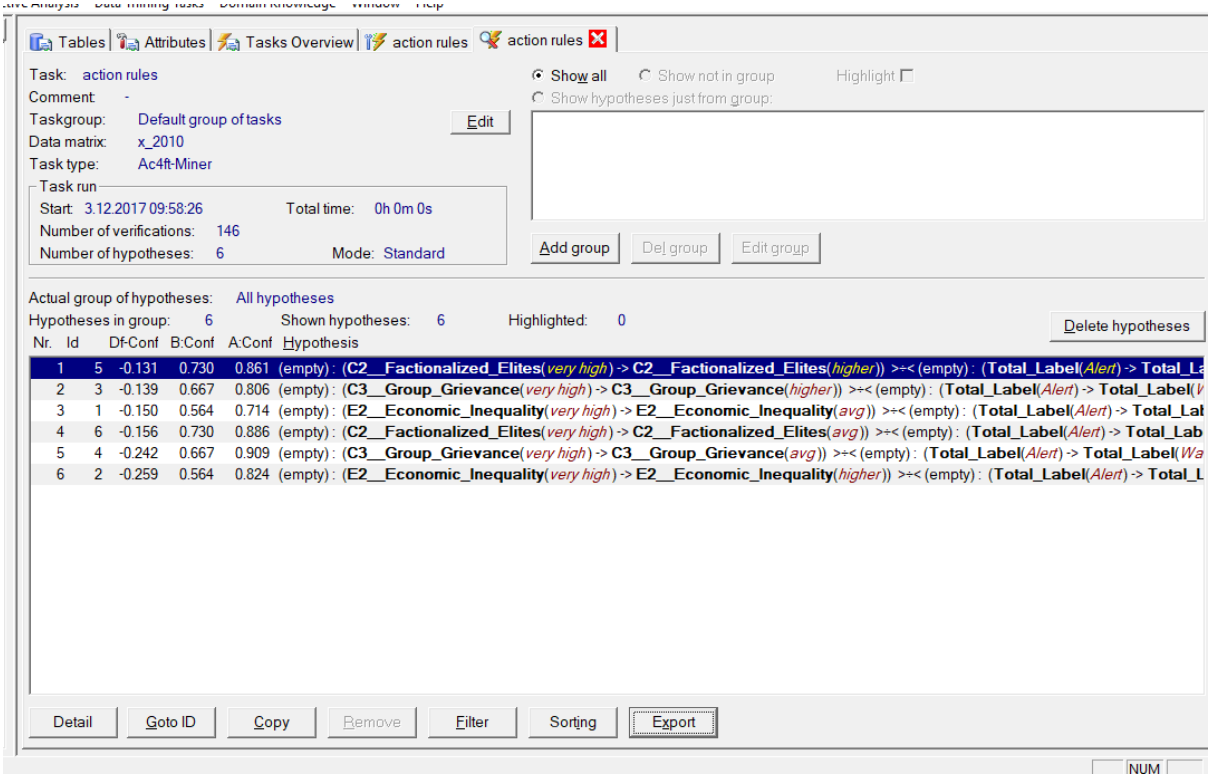


- **Action Rules for Alert to Sustainable:**
  1. (C2__Factionalized_Elites(very high) & C3__Group_Grievance(very high) -> C2__Factionalized_Elites(very low) & C3__Group_Grievance(very low)) >÷< (Total_Label(Alert) -> Total_Label(Sustainable))

2. (C3__Group_Grievance(very high) -> C3__Group_Grievance(very low)) >÷< (Total_Label(Alert) -> Total_Label(Sustainable))

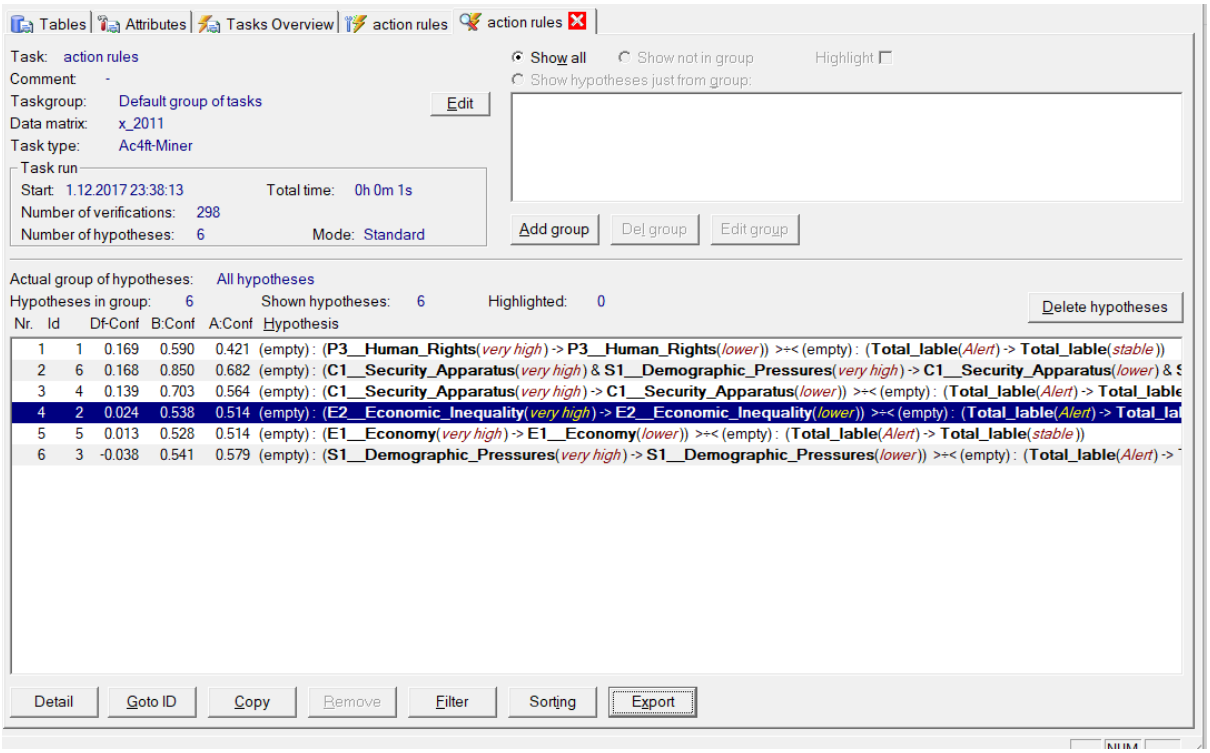


- **Action Rules for Alert to Warning:**
  1. (C2__Factionalized_Elites(very high) -> C2__Factionalized_Elites(higher)) >÷< (Total_Label(Alert) -> Total_Label(Warning))
  2. (C3__Group_Grievance(very high) -> C3__Group_Grievance(higher)) >÷< (Total_Label(Alert) -> Total_Label(Warning))

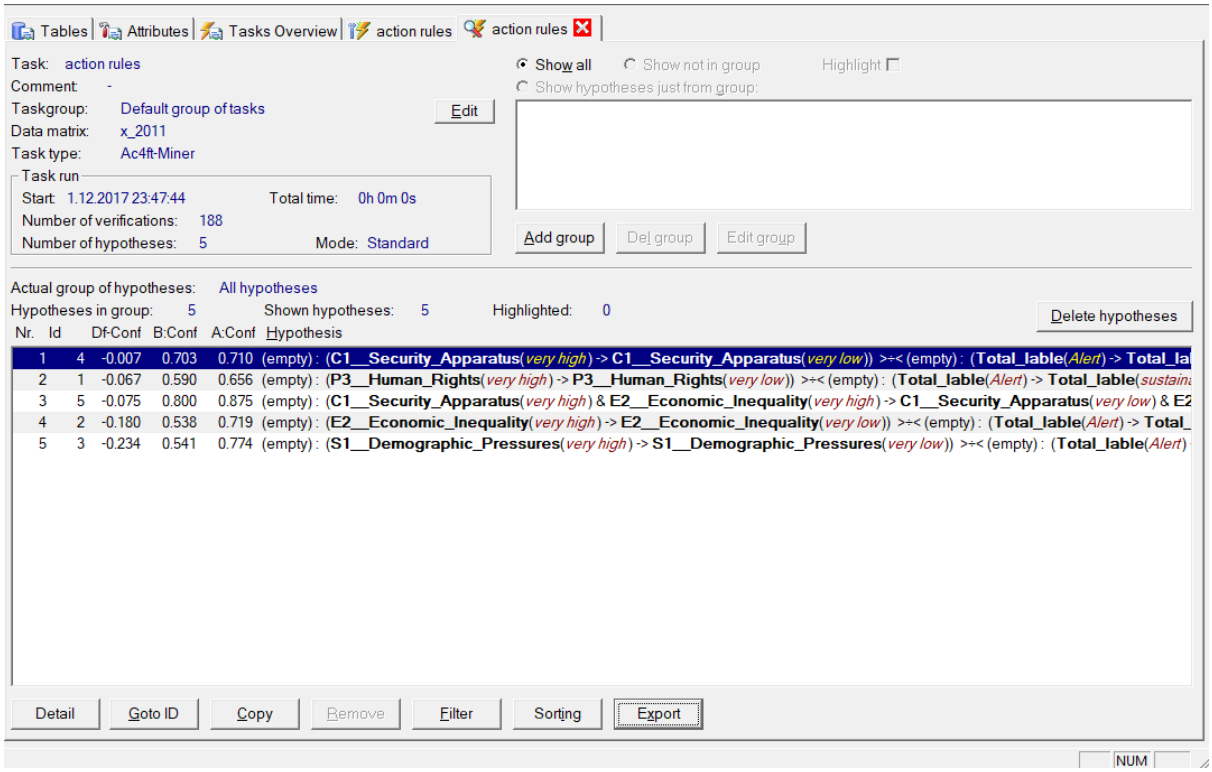

Tables
Attributes
Tasks Overview
action rules
action rules
Task: action rules
Comment: -
Taskgroup: Default group of tasks
Data matrix: x_2010
Task type: Ac4ft-Miner
Edit
Task run
Start: 3.12.2017 09:58:26
Total time: 0h 0m 0s
Number of verifications: 146
Number of hypotheses: 6
Mode: Standard
Show all
Show not in group
Highlight
Show hypotheses just from group:
Add group
Del group
Edit group
Actual group of hypotheses: All hypotheses
Hypotheses in group: 6
Shown hypotheses: 6
Highlighted: 0
Delete hypotheses
Nr. Id Df-Conf B:Conf A:Conf Hypothesis
1 5 -0.131 0.730 0.861 (empty) : (C2__Factionalized_Elites(very high) -> C2__Factionalized_Elites(higher)) >+< (empty) : (Total_Label(Alert) -> Total_La
2 3 -0.139 0.667 0.806 (empty) : (C3__Group_Grievance(very high) -> C3__Group_Grievance(higher)) >+< (empty) : (Total_Label(Alert) -> Total_Label(
3 1 -0.150 0.564 0.714 (empty) : (E2__Economic_Inequality(very high) -> E2__Economic_Inequality(avg)) >+< (empty) : (Total_Label(Alert) -> Total_Lab
4 6 -0.156 0.730 0.886 (empty) : (C2__Factionalized_Elites(very high) -> C2__Factionalized_Elites(avg)) >+< (empty) : (Total_Label(Alert) -> Total_Lab
5 4 -0.242 0.667 0.909 (empty) : (C3__Group_Grievance(very high) -> C3__Group_Grievance(avg)) >+< (empty) : (Total_Label(Alert) -> Total_Label(Wa
6 2 -0.259 0.564 0.824 (empty) : (E2__Economic_Inequality(very high) -> E2__Economic_Inequality(higher)) >+< (empty) : (Total_Label(Alert) -> Total_L
Detail
Goto ID
Copy
Remove
Filter
Sorting
Export
NUM

For Year 2011:

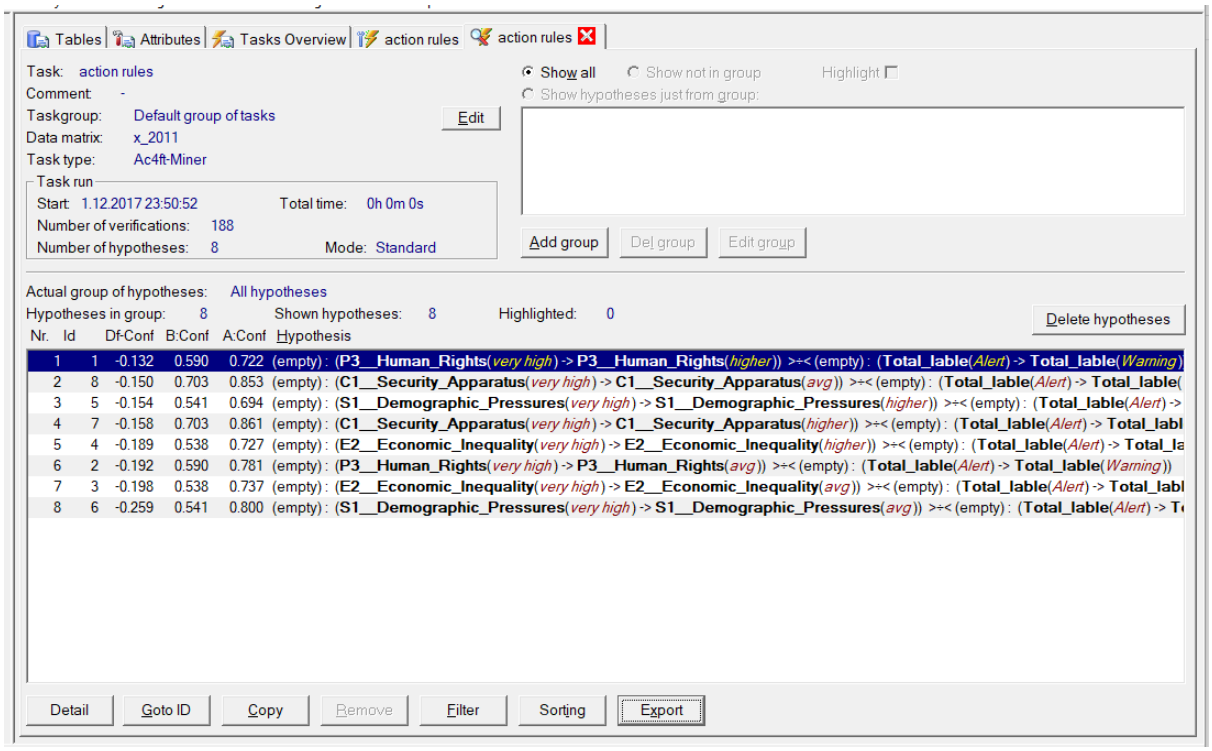
- **Action Rules for Alert to Stable:**
  1. (E2__Economic_Inequality(very high) -> E2__Economic_Inequality(lower)) >÷< (Total_lable(Alert) -> Total_lable(stable))
  2. (S1__Demographic_Pressures(very high) -> S1__Demographic_Pressures(lower)) >÷< (Total_lable(Alert) -> Total_lable(stable))



- **Action Rules for Alert to Sustainable:**
  1. (E2__Economic_Inequality(very high) -> E2__Economic_Inequality(very low)) >÷< (Total_lable(Alert) -> Total_lable(sustainable))
  2. (S1__Demographic_Pressures(very high) -> S1__Demographic_Pressures(very low)) >÷< (Total_lable(Alert) -> Total_lable(sustainable))

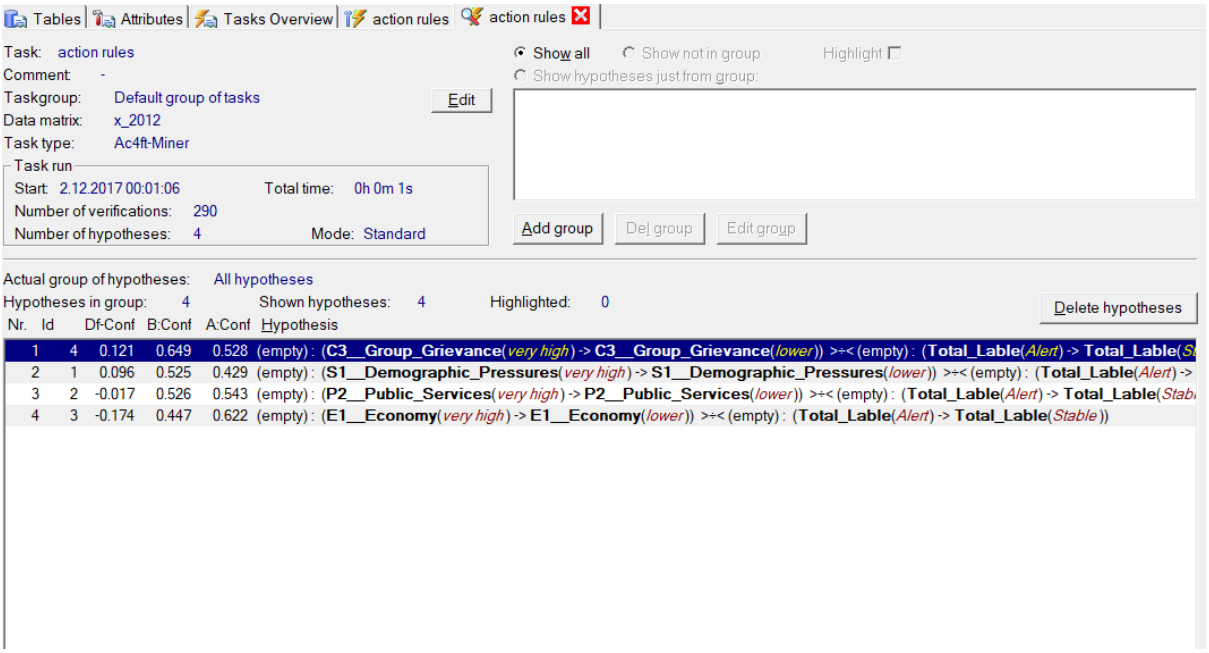


- **Action Rules for Alert to Warning:**
  1. (P3__Human_Rights(very high) -> P3__Human_Rights(higher)) >÷< (Total_lable(Alert) -> Total_lable(Warning))
  2. (C1__Security_Apparatus(very high) -> C1__Security_Apparatus(avg)) >÷< (Total_lable(Alert) -> Total_lable(Warning))

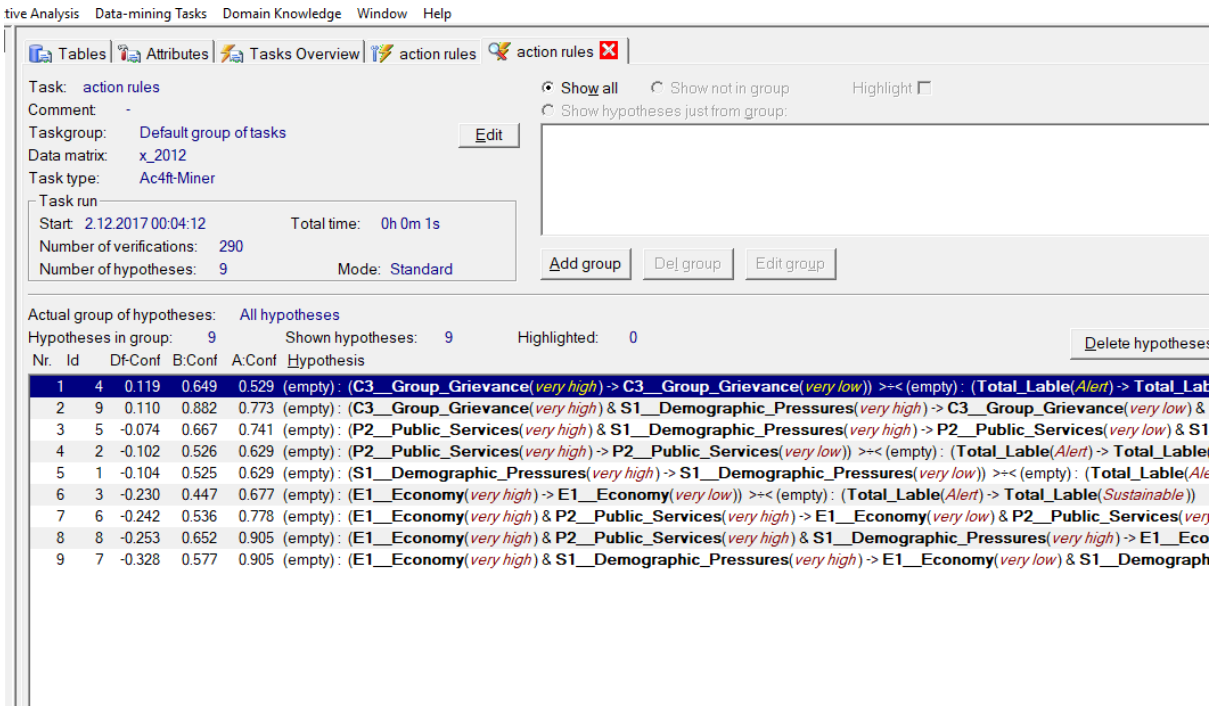


For Year 2012:

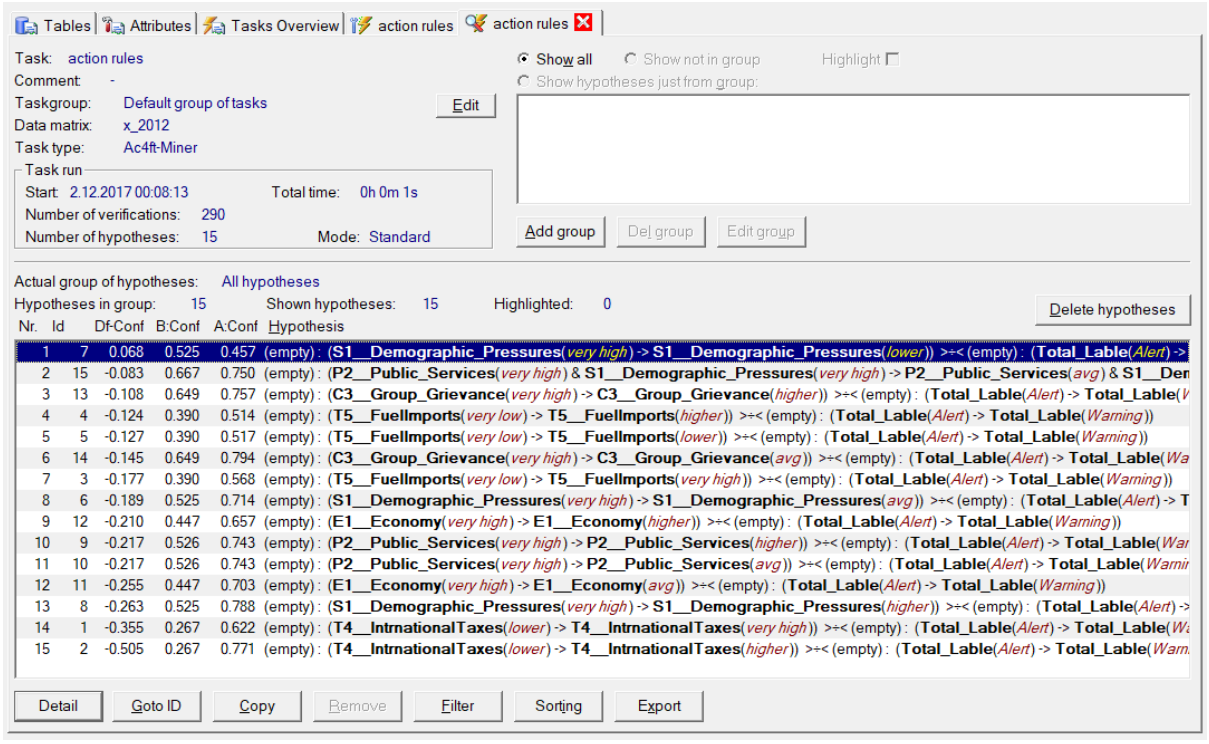
- **Action Rules for Alert to Stable:**
  1. (C3__Group_Grievance(very high) -> C3__Group_Grievance(lower)) >÷< (Total_Lable(Alert) -> Total_Lable(Stable))
  2. (S1__Demographic_Pressures(very high) -> S1__Demographic_Pressures(lower)) >÷< (Total_Lable(Alert) -> Total_Lable(Stable))



- **Action Rules for Alert to Sustainable:**
  1. (C3__Group_Grievance(very high) & S1__Demographic_Pressures(very high) -> C3__Group_Grievance(very low) & S1__Demographic_Pressures(very low)) >÷< (Total_Lable(Alert) -> Total_Lable(Sustainable))

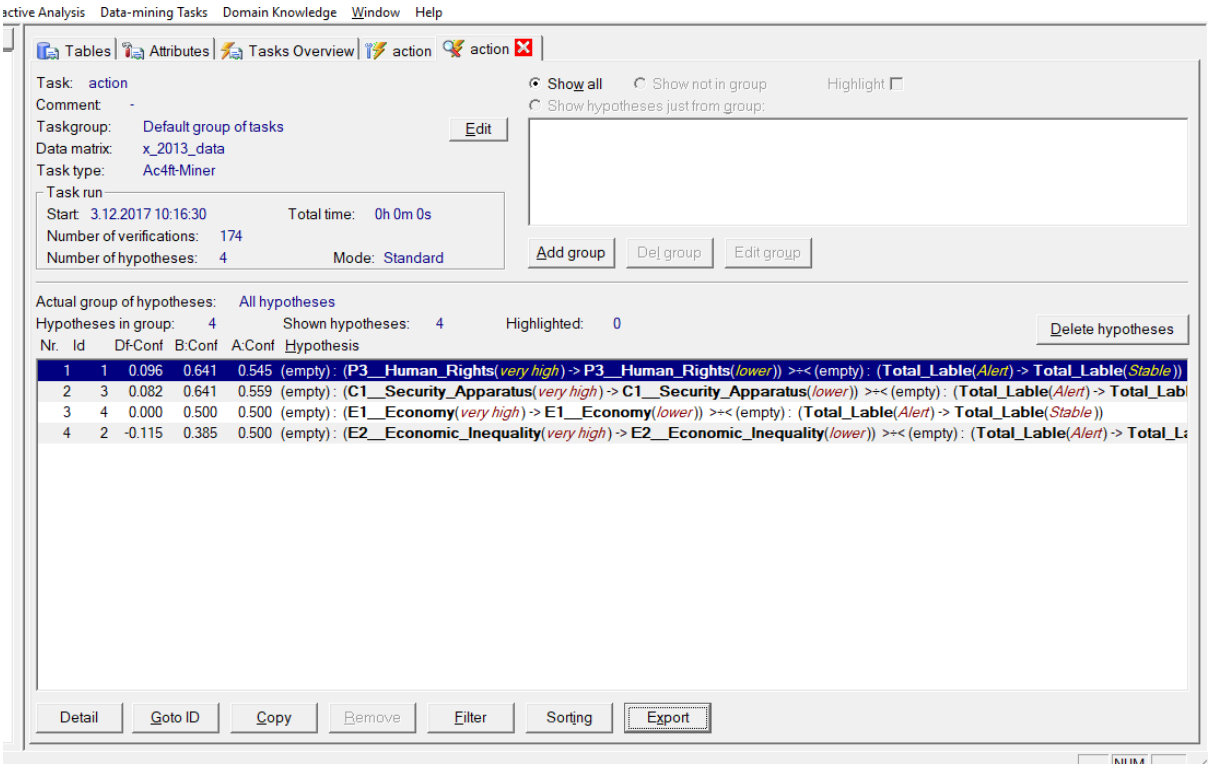


- **Action Rules for Alert to Warning:**
  1. (T5__FuelImports(very low) -> T5__FuelImports(higher)) >÷< (Total_Lable(Alert) -> Total_Lable(Warning))
  2. (P2__Public_Services(very high) -> P2__Public_Services(higher)) >÷< (Total_Lable(Alert) -> Total_Lable(Warning))

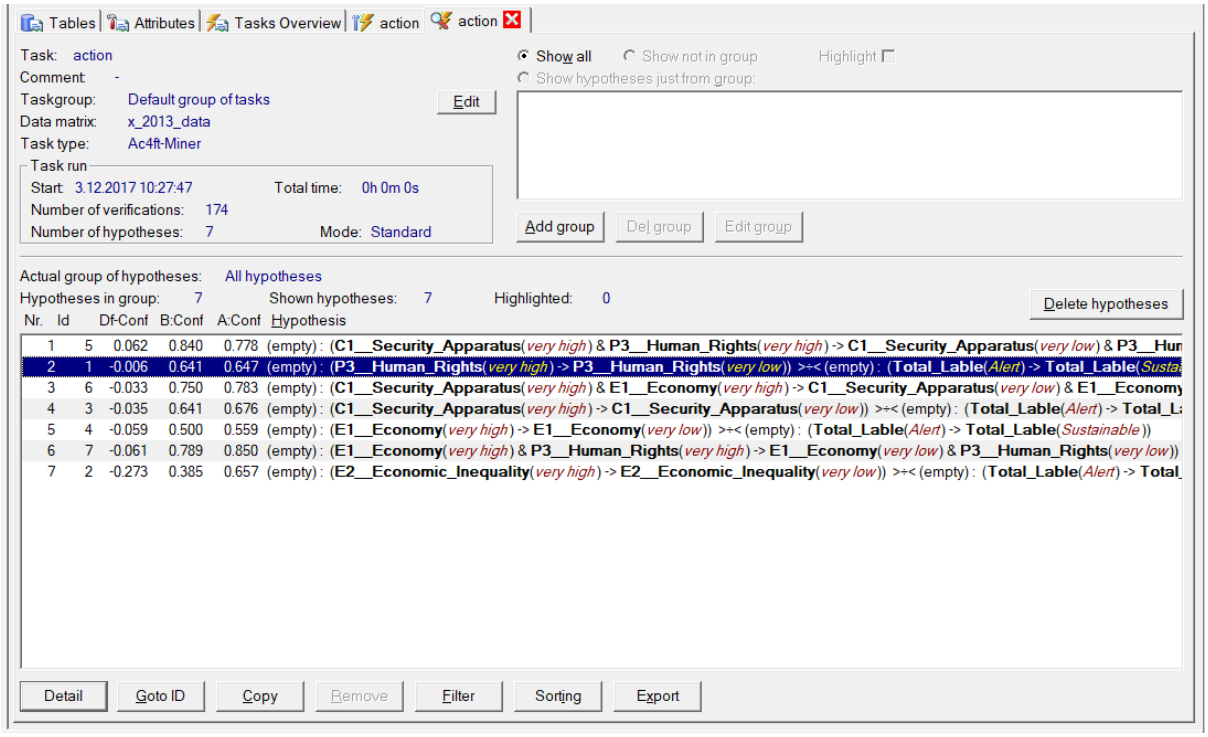


For Year 2013:

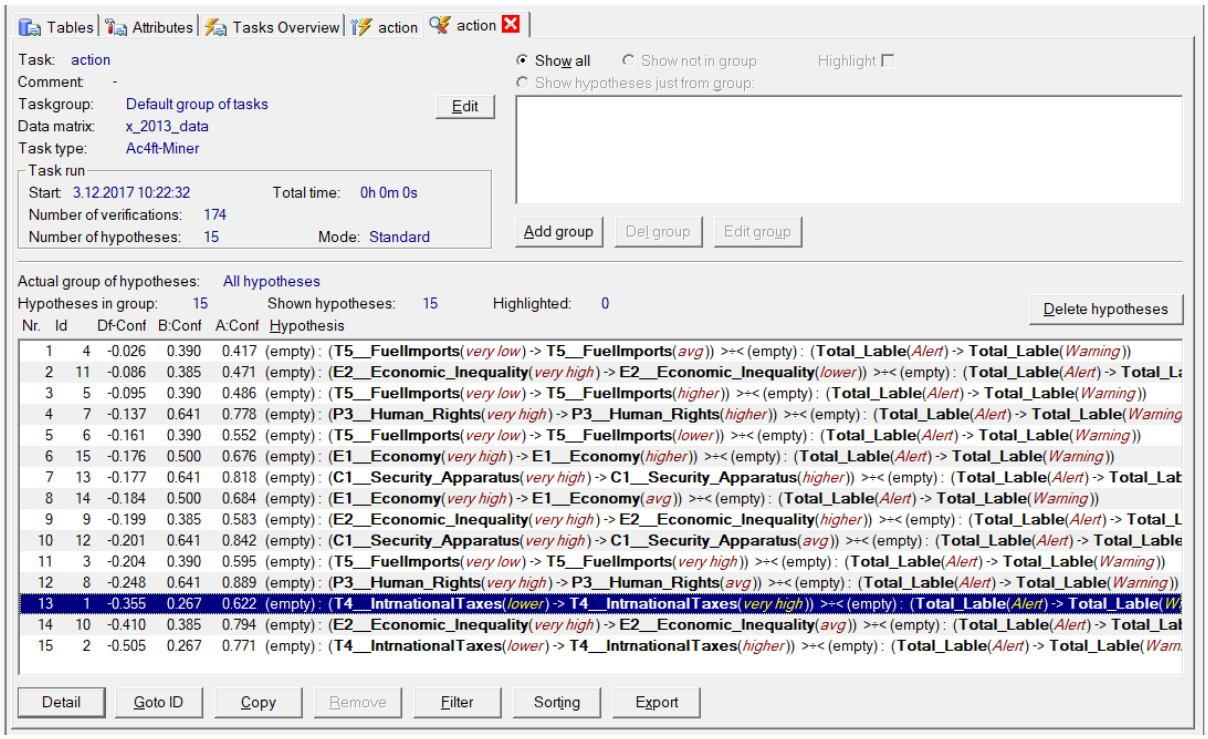
- **Action Rules for Alert to Stable:**
  1. (E2__Economic_Inequality(very high) -> E2__Economic_Inequality(lower)) >÷< (Total_Lable(Alert) -> Total_Lable(Stable))
  2. (P3__Human_Rights(very high) -> P3__Human_Rights(lower)) >÷< (Total_Lable(Alert) -> Total_Lable(Stable))



- **Action Rules for Alert to Sustainable:**
  1. (C1__Security_Apparatus(very high) & P3__Human_Rights(very high) -> C1__Security_Apparatus(very low) & P3__Human_Rights(very low)) >÷< (Total_Lable(Alert) -> Total_Lable(Sustainable))
  2. (P3__Human_Rights(very high) -> P3__Human_Rights(very low)) >÷< (Total_Lable(Alert) -> Total_Lable(Sustainable))



- **Action Rules for Alert to Warning:**
  1. (T5__FuelImports(very low) -> T5__FuelImports(avg)) >÷< (Total_Lable(Alert) -> Total_Lable(Warning))
  2. (T4__IntrnationalTaxes(lower) -> T4__IntrnationalTaxes(very high)) >÷< (Total_Lable(Alert) -> Total_Lable(Warning))



REFERENCES

- http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators#
  For downloading the data of countries based on World Development Indicators
- http://fundforpeace.org/fsi/
  For getting the data for four years containing 12 World Development Indicators to determine the Alertness of world's countries.
- http://visionofhumanity.org/about/ and https://en.wikipedia.org/wiki/Global_Peace_Index
  For getting the information related to Global Peace Indicator
- http://www.who.int/gho/health_financing/total_expenditure/en/
  https://en.wikipedia.org/wiki/List_of_countries_by_total_health_expenditure_per_capita
  For getting the information about the Health Expenditure total (%GDP).
- https://en.wikipedia.org/wiki/List_of_countries_by_life_expectancy
  Life expectancy at birth
- https://en.wikipedia.org/wiki/World_energy_consumption
  Fossil fuel energy consumption (% of total)
- https://www.cs.waikato.ac.nz/ml/weka/
  Weka Tool
- http://www.datapreparator.com/
  Data Preparator Software.

APPENDIX

Attribute Information