FRAGILE STATE INDEX ANALYSIS

USING WEKA AND LISPMINER

REPORT PREPARED BY:

VISHAK LAKSHMAN SANJEEVIKANI MURUGESH

(800985356)

Introduction

The Fragile States Index(FSI) is an annual report published by the Fund for Peace and the American magazine Foreign Policy since 2005. The list aims to assess states and countries based on their vulnerability to conflict or collapse and ranking all sovereign states with membership in the United Nations based on this index. This index defines if the state is susceptible to social and economic meltdown or not and helps the UN to take extra efforts in the areas ranked very high.

The index uses 12 parameters that as categorified as social (four), economic (two) and political (six). The social indicators include demographic pressures, refugees and internally displaced persons, group grievance and human flight and brain drain. The economic indicators are uneven economic development, poverty and economic decline. The political factors that contribute are state legitimacy, public services, human rights and Law, security apparatus, factionalized elites and other external intervention.

Fragile State Index has been categorized into 4 different levels to label the alertness and socio-economic stability of the countries:

- Alert
- Warning
- Stable
- Sustainable

These categories have been uniformly scaled in a 0-120 scale and each factor on a scale 0-10 with 10 being the highly alert condition.

Problem Description

The objective of this project is to add six more new features that has direct impact in calculating the FSI and recalculating the total. This new total is used as a base for running the different classifying algorithms using WEKA, and determine the best classifier suited for this project by comparing their precision and accuracy. Once the classifier has been determined, we aim to extract action rules using LISp Miner that hypothesize various solutions to move a country from a state of alert to less dangerous state. We have chosen a sequence of datasets from 2012-2015 for our analysis.

Added Features

We aim to approximate the alertness of each countries by adding six new features as parameters to estimate the new total for FSI. The six new features that have identified to have a substantial impact are listed as below:

1. Expense (% of GDP)

The need to consider expense of a country in terms of percentage of GDP is of utmost importance. This gives us an idea of what the expenses of the country are of the total production value it is capable of. This has direct effect of the fragility of the state wherein the financial and infrastructure are direct beneficiaries.

2. Inflation in Consumer Goods (% of GDP)

Inflation in consumer goods gives us an idea of the expense of living in the country and how the much consumption is occurring despite it. Inflation has a large effect on the economy be it production or sale and how much resources are being used for consumption. For Example, one of the main factors influencing demand for consumer goods is the level of employment. The more people there are receiving a steady income, the more people there are who are able to make discretionary spending purchases.

3. Exports of goods and services (% of GDP)

Exports are an essential part of the country which have the need for the use of exporting and hence have a high impact on the FSI. The amount of exports again helps in estimating the dependency of a country on another. The exports also show how a country can manage to create income, jobs and necessary infrastructure for exporting. Considering these side factors and impact it is necessary to include this as a parameter for the FSI.

4. Unemployment, total (% of total labor force) (modeled ILO estimate)

Unemployment is forever an issue, this usually is a result of higher population and lower opportunities in the job sector of a country. This could also could mean inability of a country to create jobs for it's citizens and has long term affects.

5. Health expenditure, public (% of GDP)

Health is a paramount factor to determine the stability of the country. People are assets to any country and the health is important. Although low expenditure could mean the country does not intent to value quality life of the citizens, higher expenditure can also mean prevalent of some disease in the country and that is not a good sign for the country and should have effect on the FSI.

6. Renewable Energy Production (% of total energy)

Renewable energy is of utmost importance to every country to satisfy its energy needs be it in commercial or non-commercial consumption. Non-renewable energy will someday be non-existent and can have adverse on the country if renewable energy sector is not invested to make the maximum of natural resources as solar energy, wind energy etc, therefore not required to depend fully on natural oil or gas.

Data Gathering and Preparation

The initial data has been procured from http://fsi.fundforpeace.org/ involving initial 12 features. The rest of the 6 new features were procured. The initial data was cleaned and converted by labelling using the four tags of ALERT, WARNING, STABLE & SUSTAINABLE. The new features were a part of the list of all 265 countries, so we used VLOOKUP to merge the yearly add with their corresponding new features. Once all the necessary countries were taken into consideration after matching corresponding feature, further work was commenced. File conversions from excel were made into CSVs, used by LISpMiner, and ARFF data files for Weka.

Initial Data Analysis

This was our first experience with Weka, so to get a grip of the tool, we tried running sample classifiers on dataset pertaining to year 2012 and obtained the following summary for the different algorithms.

J48

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	152	85.3933 %
Incorrectly Classified Instances	26	14.6067 %
Kappa statistic	0.7733	
Mean absolute error	0.0797	
Root mean squared error	0.2623	
Relative absolute error	24.678 %	
Root relative squared error	65.3983 %	
Total Number of Instances	178	

LMT

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	165	92.6966 %
Incorrectly Classified Instances	13	7.3034 %
Kappa statistic	0.8844	
Mean absolute error	0.0483	
Root mean squared error	0.1571	
Relative absolute error	14.9692 %	
Root relative squared error	39.1516 %	
Total Number of Instances	178	

NaiveBayes

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	172	96.6292 %
Incorrectly Classified Instances	6	3.3708 %
Kappa statistic	0.9473	
Mean absolute error	0.0206	
Root mean squared error	0.1199	
Relative absolute error	6.3866 %	
Root relative squared error	29.8915 %	
Total Number of Instances	178	

KStar

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	161	90.4494 %
Incorrectly Classified Instances	17	9.5506 %
Kappa statistic	0.848	
Mean absolute error	0.0557	
Root mean squared error	0.2021	
Relative absolute error	17.2607 %	
Root relative squared error	50.3942 %	
Total Number of Instances	178	

From the above summary, we can infer that NaiveBayes tend to have the highest accuracy in classifying instances and LMT comes the closest second. Hence, we decided to implement these two algorithms for classifying datasets including the six features we have

added. Before adding our features, we have tabulated the relative precision of all the initial 12 features present in the dataset.

Feature	Correctly Classified Instances (%)
C1. Security Apparatus	64.3713 %
C2. Factionalized Elites	63.5262 %
C3. Group Grievance	72.8805 %
E1. Economy	69.7762 %
E2. Economic Inequality	68.3704 %
E3. Human Flight and Brain Drain	73.2341 %
P1. State Legitimacy	63.6231 %
P2. Public Services	65.2596 %
P3. Human Rights	65.0169 %
S1. Demographic Pressures	67.2506 %
S2. Refugees.and IDPs	72.0417 %
X1. External Intervention	70.1536 %

Post Addition Analysis:

NaiveBayes:

The classification technique based on Bayes' Theorem which assumes the predictors are not dependant on each other. So, when we run this classifier what happens is that all the 18 features are considered not related to each other and provides a one-dimensional distribution of how the countries are classified with higher precision and at a faster rate.

LMT:

A Logistic Model Tree (LMT) is a supervised training algorithm that combines logistic regression (LR) and decision tree learning. Building an LMT takes place in three steps: First, a decision tree is built so that it has linear regression models at its leaves. This provides a piecewise linear

regression model. The next step is to implement a LogitBoost algorithm at every node in the tree so that it produces a Logistic Regression model at every node. The C4.5 algorithm is then used split each of the node from its tree and pruned giving the desired end product. The various classifier results for years 2012-2015 are given below:

Year 2012

	NaiveBayes	LMT	KStar
Precision	89.8876 %	90.4494 %	87.6404 %
Correctly Classified Instances	160	161	156
Kappa statistic	0.8259	0.8259	0.7778
Mean absolute error	0.0539	0.0539	0.0596
Relative absolute error	18.8282 %	20.3904 %	20.805 %

Year 2013

	NaiveBayes	LMT	KStar
Precision	83.7079 %	90.4494 %	87.0787 %
Correctly Classified Instances	149	161	155
Kappa statistic	0.6975	0.7991	0.7363
Mean absolute error	0.0818	0.0583	0.0652
Relative absolute error	32.8287 %	23.384 %	26.19 %

Year 2014

	NaiveBayes	LMT	KStar

Precision	88.764 %	92.1348 %	89.8876 %
Correctly Classified Instances	158	164	160
Kappa statistic	0.8105	0.8665	0.8205
Mean absolute error	0.055	0.0506	0.051
Relative absolute error	18.619 %	17.1237 %	17.2715 %

Year 2015

	NaiveBayes	LMT	KStar
Precision	87.6404 %	87.0787 %	87.6404 %
Correctly Classified Instances	156	155	156
Kappa statistic	0.7967	0.7801	0.7864
Mean absolute error	0.0615	0.0731	0.0672
Relative absolute error	20.6656 %	24.5637 %	22.5676 %

From the above observations, it is evident that LMT has the best average precision. KStar classifier comes the close second. The detailed view of LMT classifier for year 2012 is given below:

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 161 90.4494 %

Incorrectly Classified Instances 17 9.5506 %

Kappa statistic 0.8295

Mean absolute error 0.0584

Root mean squared error	0.2022
Relative absolute error	20.3904 %
Root relative squared error	53.6326 %
Total Number of Instances	178

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.862	0.020	0.893	0.862	0.877	0.854	0.985	0.934	Alert
	0.900	0.043	0.857	0.900	0.878	0.842	0.983	0.931	Stable
	0.000	0.000	0.000	0.000	0.000	0.000	0.983	0.750	Sustainable
	0.943	0.111	0.926	0.943	0.935	0.836	0.972	0.982	Warning
Weighted Avg.	0.904	0.079	0.889	0.904	0.897	0.826	0.977	0.959	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
25	0	0	4	a = Alert
0	36	0	4	b = Stable
0	0	3	0	c = Sustainable
3	3	0	100	d = Warning

Result

Though LMT can consume a considerate amount of time than other classifiers, it still provides a better precision. The reason is because the decision states only involve four of Alert, Stable, Sustainable and Warning and a multi-nominal LMT being the best suited for such a dataset, we can get an average 90% precision.

Action Rules

Once we are done with the classifiers, we used LISpMiner for deriving the action rules that hypothesize the various solutions that could move a country from a state of Alert to a state of less certainty. We considered the actions rules for the following three transitions:

- Alert -> Warning
- Alert -> Stable
- Alert -> Sustainable

After analysis of all years from 2012-2015, some of the most prominent rules with high confidence are provided below. For the full hypotheses list, refer the LispMiner output folder in the input_output.zip zip file, attached along with this report.

Alert -> Warning

Action Rule	DConf	BConf	AConf
(Exports(very high) -> Exports(avg)) >÷< (Total(Alert) -> Total(Warning))	-0.0114660115	0.3319327731	0.4325842697
(C1__Security_Apparatus(very high) -> C1__Security_Apparatus(lower)) >÷< (Total(Alert) -> Total(Warning))	-0.0804511278	0.3175990676	0.404494382
(Inflation(very high) -> Inflation(lower)) >÷< (Total(Alert) -> Total(Warning))	-0.0292397661	0.2621778638	0.4943820225

Here, we can infer that the increase in export of goods and services for a country can reduce the fragility of the state by filling up the coffers of government. The confidence of 43% only states this hypothesis has a major role reducing the fragility of country from alert to warning. Similarly, we can find other rules like better security apparatus can help bring down the anarchy existing within the alert states as well the reduction in inflation of consumer goods price also has a prominent effect as it has a confidence of 49%.

Alert -> Stable

Action Rule	DConf	BConf	AConf
(Health(very high) -> Health(very low)) >÷< (Total(Alert) -> Total(Stable))	0.075	0.1146245059	0.0561797753
(C1__Security_Apparatus(very high) & P1__State_Legitimacy(very high) -> C1__Security_Apparatus(very low) & P1__State_Legitimacy(very low)) >÷< (Total(Alert) -> Total(Stable))			
(Inflation(very high) -> Inflation(very low)) >÷< (Total(Alert) -> Total(Stable))	-0.0895752896	0.0142140468	0.0224719101

For reducing a state from alert -> stable, we can see that the increase in health services can make a great deal of change. Inflation of consumer goods prices though has it own role to play, it's lower confidence can say its impact isn't as vigorous as other features but still it has its own role to play. Finally, a well enforced law system and security forces can bring down the seriousness of the fragility to stable.

Alert -> Sustainable

Action Rule	DConf	BConf	AConf
(C1__Security_Apparatus(very high) & P1__State_Legitimacy(very high) & P3__Human_Rights(very high) -> C1__Security_Apparatus(very low) & P1__State_Legitimacy(very low) & P3__Human_Rights(very low)) >÷< (Total(Alert) -> Total(Sustainable))	0.0177133655	0.0328282828	0.0280898876
(HealthExpense(very high) & RenewableEnergy(very high) -> HealthExpense(very low) & RenewableEnergy(very low)) >÷< (Total(Alert) -> Total(Sustainable))	-0.0205128205	0.0036855037	0.0168539326
(RenewableEnergy(very high) -> RenewableEnergy(very low)) >÷< (Total(Alert) -> Total(Sustainable))	-0.1593984962	-0.0784313725	-0.0224719101

To make the most ambitious of changes, the following rules, hypothesize the necessary changes that had to brought in order to make an alert country into a sustainable one. Features like law system, security forces and health systems are emphasised again.

Additional to them, the increase in production and usage of renewable energy also plays it part.

Conclusion

Having ventured into softwares like Weka and LISpMiner for the first time, it took us while for us to get of hold of how to get results and look for the expected results. There were some challenges in preparing and cleaning the dataset from the crude source, but once we had a clean data set, the mining processes involing LISpMiner and the classifying using Weka helped us in achieving a strong foundation for action rules mining and the corresponding knowledge discovery from those data sets.

REFERENCES

1. http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators#
2. https://en.wikipedia.org/wiki/Fragile_States_Index
3. http://fsi.fundforpeace.org/
4. http://foreignpolicy.com/fragile-states-index-2016-brexit-syria-refugee-europe-anti-migrant-boko-haram/
5. https://www.newsecuritybeat.org/2012/06/what-are-the-most-important-factors-in-the-failed-states-index/