Classification of fragile states based on machine learning

Yuquan Li¹, Hehua Yao¹

¹Department of Computer Technology and Application, Qinghai University, Xining, China

Abstract. The study of fragile states has become a significant issue in global security, development and poverty at present. The existing classification methods of fragile state, which is a simple addition to the national index and threshold segmentation, is not reasonable enough. We introduce a new method based on machine learning. With this method, it will be easier and more reasonable to classify a country. We use two kinds of classifier, one of which is the support vector machine, and the other is the gradient boosted regression trees. Both models have flaws, so we use ensemble learning techniques to combine them. First of all, subjective labelling of a part of the national data to allows the machine to learn why a country becomes vulnerable from these data, and how to classify the vulnerability class of a country. Then, we trained the model with the data, and divided fragile states into four categories successfully (Alert, Warning, Stable and Sustainable). For the classification result, our model got a 93% test error rate, and a 96% training error rate, which is better than 77% with the threshold segmentation method.

1 Introduction

At present the study of fragile states or of the country's vulnerability research has become an international academics and policy makers to discuss in today's global security, development and poverty problems [1~3], as one of the core issues. Understanding the vulnerability of the state, and the consequences of its consequences, has only recently emerged as a field of study [4]. For the past five years, how to classify a fragile state has become an international group's main priorities [5~6]. Some well-known universities and other research institutes in many countries have set up research centers to discuss the issue of "fragile states", and academic papers are constantly being rolled out for the grand view [7~9].

In the past, fragile states were simply separated by subjective thresholds [10~11]. First, the staff evaluates economic, environmental and other indicators (scores from 0 to 10). Then, they simply add up and choose thresholds according to some rules, separating different countries using these thresholds.

In fact, it is not fair or reasonable for some countries to do so. A typical example: a country with 11 indicators is perfect, but one of them, human rights, is very low with only two points. At this time, the human rights indicator will directly affect the country's vulnerability directly, and can affect the country's vulnerability through indirect influence (such as economic inequality, group grievance). This country is supposed to be classified as endangered, but it is recognized as excellent by threshold segmentation method.

2 Support Vector Machine

SVM is a new generation machine learning method proposed by Vapnik on the basis of statistical learning theory [12]. It has significant advantages in solving small sample, nonlinear, high dimension and other problems. At present SVM has been widely used in many fields of classification problems and regression problems, and has a good prediction effect.

SVC is an abbreviation of support vector classification and is an important branch of support vector machine (SVM) [13].

The classification schematic diagram is as follows:

Figure1. SVC schematic diagram

Given training vectors $x_i \in \mathbb{R}^p$, $i=1, \dots, n$, in two classes, and a vector $y \in \{1, -1\}^n$, SVC solves the following primal problem:

$$\min_{\omega, b, \zeta} \frac{1}{2} \omega^T \omega + C \sum_{i=1}^n \zeta_i \quad (1)$$

Subject to

$$y_i(\omega^T \phi(x_i) + b) \geq 1 - \zeta_i$$

$$\zeta_i \geq 0, i = 1 \dots n \quad (2)$$

Its dual is

$$\min_{\alpha} \frac{1}{2} \alpha^T Q - e^T \alpha \quad (3)$$

Subject to

$$y^T \alpha = 0$$

$$0 \leq \alpha_i \leq C, i = 1 \dots n \quad (4)$$

where $\mathbf{e}$ is the vector of all ones, $C > 0$ is the upper bound, Q is an n by n positive semidefinite matrix, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$, where $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel. Here training vectors are implicitly mapped into a higher (maybe infinite) dimensional space by the function ϕ .

The decision function is:

$$\text{sgn}\left(\sum_{i=1}^n y_i \alpha_i K(x_i, x) + \rho\right) \quad (5)$$

3 Gradient Boosted Regression Trees

Gradient Boosted Regression Trees is a generalization of boosting to arbitrary differentiable loss functions [14~16]. GBRT is an accurate and effective off-the-shelf procedure that can be used for both regression and classification problems. Gradient Tree Boosting models are used in a variety of areas including Web search ranking and ecology [16].

The classification schematic diagram is as follows:

Figure 2. GBRT schematic

GBRT considers additive models of the following form:

$$F(x) = \sum_{m=1}^M \gamma_m h_m(x) \quad (6)$$

where $h_m(x)$ are the basis functions which are usually called weak learners in the context of boosting. Gradient Tree Boosting uses decision trees of fixed size as weak learners. Decision trees have a number of abilities that make them valuable for boosting, namely the ability to handle data of mixed type and the ability to model complex functions.

Similar to other boosting algorithms GBRT builds the additive model in a forward stage wise fashion:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (7)$$

At each stage the decision tree $h_m(x)$ is chosen to minimize the loss function L given the current model F_{m-1} and its fit $F_{m-1}(x_i)$

$$F_m(x) = F_{m-1}(x) + \underset{h}{\operatorname{argmin}} \sum_{i=1}^n L(y_i, F_{m-1}(x_i) + h(x)) \quad (8)$$

The initial model F_0 is problem specific, for least-squares regression one usually chooses the mean of the target values.

Gradient Boosting attempts to solve this minimization problem numerically via steepest descent: The steepest descent direction is the negative gradient of the loss function evaluated at the current model F_{m-1} which can be calculated for any differentiable loss function:

$$F_m(x) = F_{m-1}(x) - \gamma_m \sum_{i=1}^n \nabla_F L(y_i, F_{m-1}(x_i)) \quad (9)$$

Where the step length γ_m is chosen using line search:

$$\gamma_m = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^n LL(y_i, F_{m-1}(x_i) - \gamma \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}) \quad (10)$$

4 SVC-GBRT model

Whether it is the SVC model or the GBRT model, the expression of the single model may not be enough, so we integrate the two using integrated learning and get the final integrated learning model.

The final classification schematic diagram is as follows:

Figure3. SVC-GBRT schematic diagram

SVC-GBRT is the ensemble of SVC and GBRT. The principles of SVC and GBRT are shown in the chapter 2 and 3. Below are the training formulas for ensemble learning.

For $i = 1, 2, \dots, n$:
 $S_i = \text{Train}(S_{i-1})$
 $G_i = \text{Train}(G_i)$
 $F_i = \text{Ensemble}(S_i, G_i)$
 S : SVC model
 G : GBRT model
 F : SVC-GBRT model

We train SVC and GBRT separately, and finally use ensemble learning to combine them. We use the stacking algorithm to integrate. First, the sub-classifier is used to calculate the probability distribution of the target in each category. Then the probabilities of multiple classifiers are added and averaged. You get the final probability distribution.

5 Classification of fragile states

All the data comes from the World Bank. We went to their website and downloaded the relevant data and made a lot of analysis and processing.

We got the following data (from 2007 to 2015 of every country), and we added the variance and range indicators for some unbalanced distribution of indicators. We labelled each country year by year, using human annotation.

Table 1. Data features

No	Features
1	Security Apparatus
2	Factionalized Elites
3	Group Grievance
4	Economy
5	Economic Inequality
6	Human Flight and Brain Drain
7	State Legitimacy
8	Public Services
9	Human Rights
10	Demographic Pressures
11	Refugees and IDPs
12	External Intervention
13	Variance
14	Range

We split up the data set we previously obtained, and we got the training set, validation set, and test set for cross validation. The code was written by Python, then the model was trained through cross-validation, and the best training iterations was 600. Finally, we obtained the following classification error curve(Figure 5).

We used SVC and GBRT as the comparison, as shown in the figure 5, the test error rate of the SVC-GBRT model reached an astonishing 0.07 (while GBRT was 0.12, and the SVC was 0.14). On the other hand, the training error rate of the SVC-GBRT model reached an astonishing 0.04 (while GBRT was 0.5 and SVC was 0.14).

Figure5. Training and testing error curves

After the model training was completed, we used the model to predict a typical country, the Central African Republic, and the results were as follows (Figure 6).

As we can see from the figure 6, the probability of Africa being set for “Alert” is 0.92, the probability of Africa being set for “Warning” is 0.3, the probability of Africa being set for “Stable” is 0.1. the probability of Africa being set for “Sustainable” is 0.04. Finally, the country was classified as “Alert”.

Figure6. Classification result of the Central African Republic in 2010

Suppose that our annotation is reasonable enough, we made a comparison between machine learning classification results and previous threshold segmentation results.

Table 2. A comparison between different methods of classification

Method	Error rate	Accuracy
Machine Learning	7%	93%
Threshold Segmentation	23%	77%

6 CONCLUSIONS

Obviously, the traditional threshold segmentation method does not classify the vulnerability class of a country reasonably and fairly. Therefore, we introduce a new classification method, using machine learning method to classify the vulnerability class of a country. First of all, subjective labelling of a part of the national data to allow the machine to learn why a country becomes vulnerable from these data. Through model training and testing, we successfully used the model to separate the country into four types (Alert, Warning, Stable and Sustainable). And the accuracy of the test data set was 93%, which was better than 77% by the threshold segmentation method.

Fund

Qinghai university youth research fund 2016.

References

1. Naude, W. & Santos- Paulino, A. U. *Fragile States: Causes, Costs, and Responses*. Oxford: Oxford University Press, 2011: 23.
2. McLoughlin, C.. *Topic guide on fragile states*. Birmingham: University of Birmingham, UK.2012: 6-29.
3. Sanin, F. G.. *Evaluating State Performance: A Critical View of State Failure and Fragility Index*. European Journal of Development Research, 2011, 23(1): 20-42.
4. Haken, N. &, Messner, J.J. *Fragile States Index*2014. Foreign Policy (July -August)2014: 10.

5. Menkhaus, K.. State Fragility as A Wicked Problem. PRISM, 2010, 1(2): 85-100.
6. Mata, J. F. & Ziaja S.. Users' Guide on Measuring Fragility. Bonn: German Development Institute, 2009: 24/5
7. Ziaja S.. What Do Fragility Indices Measure? Z Vgl Polit Wiss, 2012, 6: 39-64.
8. Call, C.. The Fallacy of The "Failed States". Third World Quarterly, 2008, (8): 1491-1507.
9. Grävingholt, J. & Ziaja S.. State Fragility: Towards a Multi- Dimensional Empirical Typology. DIE Discussion Paper, 2012.
10. Kaplan, S.. Identify Truly Fragile States. The Washington Quarterly, 2014, (37)1: 49-63.
11. Marshall, M. & Cole, B.. Global Report 2011: Conflict, Governance and State Fragility. Foreign Policy Bulletin, 2011, 18(5).
12. ["Support-vector networks"](#), C. Cortes, V. Vapnik - Machine Learning, 20, 273-297 (1995).
13. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
14. Breiman L, Friedman J, Olshen R. Classification and regression trees. et al. . 1984
15. Friedman J H. Greedy function approximation: A gradient boosting machine. The Annals of Statistics. 2001
16. Stochastic gradient boosting[J] . Jerome H. Friedman. Computational Statistics and Data Analysis . 2002 (4)