The 23rd International Conference on Artificial Intelligence,
July 26-29, 2021, Las Vegas NV, USA

A Web-based Data Analysis Platform for Collaborative Decision Assistance

Christian Díaz-Cáez and Charles Kim
Electrical Engineering and Computer Science
Howard University
Washington, DC: USA
christian.diaz@bison.howard.edu; ckim@howard.edu

***Abstract*—Political regimes and social status are always shifting in a region. It is one of the principal reasons why its analysis is needed to decipher for military or humanitarian missions for the region. More data from heterogeneous sources and locations may improve the mission decision-making, by which more accurate and less variant prediction of social unrest threat in a local area would be possible. This paper discusses a decision assistance system with a machine reasoning algorithm built on the information entropy minimum principle,and its implementation on a web-server to accommodate data upload from collaborators. The developed web-server has the capacity to add and update a database, and the updated database is supplied to the decision assistance algorithm for decision rule generation regarding social unrest threat or regime change prediction. The web-server based decision assistance system has been evaluated using publicly available polity data on regime change.**

***Index Terms*—Machine Reasoning, Inductive Reasoning, Collaborative Platform, Information Entropy, Polity Regime.**

I. Introduction

Prediction of regime change in a country or social unrest in a local area is of great interest and challenge for many agencies in government. This led to placing the highest importance on understanding people and population. In dealing with the challenge, government agencies builds a collaborative mechanism by which the decisions, risks, and responsibilities of the challenge may be shared. Collaboration and shared decision-making among all concerned parties, government agencies, military, and civilian organizations have become extremely important.

This work discusses a collaborative decision-assist system implemented on a web-based platform that generates decision rules for predicting threats of regime change or social unrest.

The importance of people and human activity has long been emphasized. Without understanding the polity regime status and populace, predicting local threats such as insurgency is almost impossible. [1]. Collection of local population and social status and their analysis are an essential component of understanding the people and region. However, in the connected global world, local data alone is not enough because it is possible that the local status may quickly change as seemingly irrelevant events in a remote area may stir or influence the local population. Considering the global influence over locals, the data collection and interpretation has to be done in a global perspective. Data collection should be diverse and collaborative, local and global, and of surveillance and intelligence. Collaborative data collection and analysis would provide timely decision making assistance in determining threat levels of social unrest or regime change in global perspective.

In previous efforts [2], we have formulated an inductive reasoning-based decision-making algorithm based on information entropy. In this paper, we briefly discuss about the decision-making algorithm and its implementation on a web-server to apply it in determining the threat level of social unrest in a region or regime change in a country. The web-based assistance system incorporates and analyzes both the region-specific populace and the global events of human dynamics with the capability of updating its decision when additional data is added to the common database.

The brain of the decision-assistance system is built upon an inductive inference algorithm that can extract decision-making rules along with probability and margin of error of the rules. Also, the decision assistant system can prioritize decision variables so that high-value variables can be used in future trend analysis for threat prediction, anticipation, and mitigation.

This paper is organized as follows. In the next chapter, we discuss the framework and development of the web-server. Then, we detail the information entropy-based inductive reasoning in Chapter III, followed by decision rule generation in Chapter IV. In Chapter V, we discuss validation of the the decision-assistant in the web-server environment. Finally, we conclude the paper in Chapter VI.

II. Web-based collaborative Data Analysis Framework

Web-based collaborative platform is a term that is widely used to describe a design that is developed for cooperating parties belonging to different entities in different places to support collaborative activities in various applications such as production, development, marketing, and service [3].

A good web-based collaborative platform is desired to support resource aggregation and collaboration so that participating parties can see the common pattern reflected in

Financial support provided by the U. S. Army through the MSI STEM Research & Development Consortium with the project entitles "Irregular Warfare Decision Assist System for Determining Threat," (Award Number: D01 _W911SR-14-2-0001-0015)

data sets, fostering shared consensus among the collaborators. For this reason, many organizations dedicate much time and money to designing and developing platforms best suited to their needs.

In our own web-based platform development, we aim to implement a collaboration platform which enables to collect data from multiple parties of different agencies and areas and generate global decision rules in a global perspective in determining population and social status. Figure 1 illustrates an example of a single layout network.

In our case, the decision-making algorithm is deployed on the server, multiple parties represent the clients, and the database represents the common storage for local and global populace and social data. By this scheme, each client can obtain more accurate and global perspective results based on the diverse data from the collaborators.

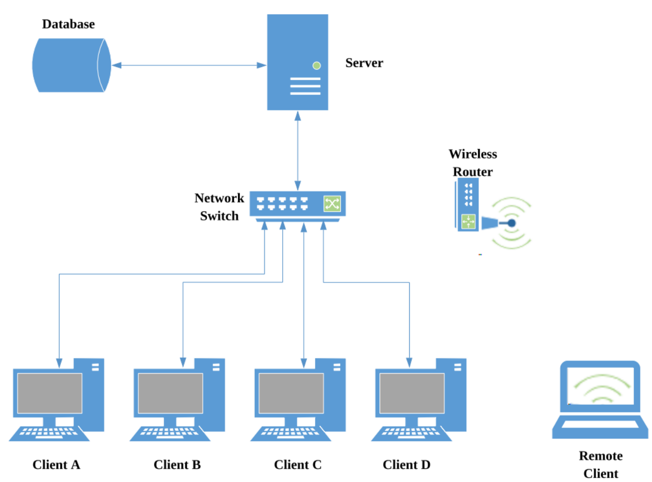

Figure 1. Example of a single layout network.

A web page is designed for the web-based platform to work as an interface through which the users will be able to upload their data and access the generated global decision rules. Apache server is used for the connection between the interface and the database.

III. Machine Reasoning for Decision Assistance

Machine reasoning entails using a computer to solve reasoning components of problems. Some questions that arise in the study are the representation of knowledge, the rules to derive new knowledge from what is available, and the strategies to control those rules. Other questions refer to implementing the resulting theory and the applications for which the corresponding software can be used [4].

In a machine reasoning algorithm, to carry out decision making using data, it looks for new facts from the data, generates conclusions from previous facts, and applies specific reasoning types known as inference. The inference we apply for the collaborative decision assistance system is information entropy and its utilization in logical induction.

A. Information Entropy

An important fact that Shannon revealed when laying the foundation of information theory was that it was possible to give the two terms, uncertainty and information, a mathematical meaning based on a probabilistic model and in terms of a numerically measurable quantity, which he called entropy. [5] [6].

The term entropy is information entropy, which is an expected value of information quantity, and thus is defined as $S = -kP \ln P$. However, the information can be related to a prior estimate of probability. Thus the quantity of information I_Q is defined as proportional to the negative of the logarithm of probability (P), and thus can be expressed as $I_Q = -k \ln P$ with k a constant corresponding to the choice of measurement unit. [7].

In drawing a decision from data, it can be said that the information measure compares the contents of the data one receives to a prior state of expectation. The higher is the prior estimate of the probability for an outcome (or a threat in our work), the lower will be the information one gains by observing its occurrence.

Therefore, entropy can be interpreted as a measure of uncertainty in an induced probability of the event. Further, the minimum entropy can be used as an ordering principle that determines which data variables used in the past decision on the event are more like those in the present decision making space. The entropy is smallest when all the information has been extracted from the available data [8], which leads to derive a decision rule optimized for the present state of information. The details of the algorithm utilizing the information entropy are in the next chapter.

B. Inductive Reasoning

Inductive reasoning is a process that involves learning by examples where a system tries to induce a general rule from a set of observed instances. This also applies a classification process assigning, to a particular input, a class to which it belongs. Inductive rule extraction and update are the principal technical approaches for the previously proposed framework [9] for the decision-assist system and threat prediction.

An aspect of a desirable decision assist system is that it has something capable of adjusting to its dynamic and varying human environment. To make such an adjustment, the system must continually be classifying different data variables and their values as equivalent or not equivalent. It has been shown that inductive reasoning with entropy minimum principle can be used as a rule extractor and learner to accommodate the changing values of the variables [9].

Inductive inference or induction uses specific observations to make broad generalizations, which conclude from the data. The three laws of induction are as follows:

- **Law 1**: Given a set of irreducible outcomes of classification, the induced probabilities are those probabilities consistent with all available information that maximize the entropy of the set.

- **Law 2**: The induced probability for the decision being a certain threat based on the overall data entries in the collected database is proportional to the probability density for the threat of the induced probability of a single data variable entry.
- **Law 3**: The induced rule (or set of rules) for a threat is that rule consistent with all available data variables for which the entropy is minimum.

The relatively unbiased probability in the equation 1 is defined from law 2, reflecting no more information from the data entries [7], as conclusively expressed below. In the equation 1, n is the total number of data entries, N is the total number of threats considered in social unrest or insurgency, and x is the number of data entries that resulted in the threat.

$$P = \frac{x+1}{n+N} \tag{1}$$

On the other hand, the third law of induction is the recognition or classification of patterns. Entropy on N number of threat classes can be defined by equation 2 [10].

$$S = k \sum_{k=1}^{N} P^{T_k} \ln P^{T_k} \tag{2}$$

IV. ENTROPY-BASED DECISION RULE GENERATION

A rule derivation in threat classification assumes that the data variables are in a binary system. Problem occurs when the data is in a continuous or numerical system. A certain value must be set by which sample values are divided into a binary system called a threshold value.

To convert the data set into a binary system, the minimum entropy principle is used to set a threshold value and separate it into two discrete values. All the values above the threshold are replaced by '1', and '0' for those below the threshold. This kind of conversion ("binarization") is done for each individual data; therefore any data received would become binary data. The threshold value determination process is explained next.

A. Threshold Value Determination

As stated above in Chapter III, since in the minimum entropy state all the information has been extracted from the available data set, a separation rule must be a screening rule which assures the probability in the equation 2 is closer to 1.0 or 0.0, in which the entropy will be the smallest.

If we want to find the threshold value for sample values of a data variable in the range $[X_1, X_2]$, where $X_2 > X_1$, an entropy equation for a certain value x as a candidate threshold [11] is presented in the equation 3.

$$S(x) = p(x)S_p(x) + q(x)S_q(x) \tag{3}$$

$$S_p(x) = -\sum_{m=1}^{N} p_m(x) \ln[p_m(x)] \tag{4}$$

$$S_q(x) = -\sum_{m=1}^{N} q_m(x) \ln[q_m(x)] \tag{5}$$

$S(x)$ is the entropy for a certain value $x > 0$, $S_p(x)$ is the entropy for the sample values in the region $[X_1, X_1 + x]$, and $S_q(x)$ is the entropy for the sample values in the region $[X_1 + x, X_2]$.

Further, $p_m(x)$ is the probability of the m^{th} outcome (or threat) class given that the variable value is in the region $[X_1, X_1 + x]$, while $q_m(x)$ is the probability of the m^{th} outcome (or threat) class given that the variable value is in the region $[X_1 + x, X_2]$ as defined below.

$$p_m(x) = \frac{n_m(x)}{n(x)} \tag{6}$$

$$p(x) = \frac{n(x)}{n} \tag{7}$$

In equations 6 and 7, $n_m(x)$ is the number of the m^{th} outcome class samples located in the range $[X_1, X_1 + x]$, $n(x)$ the total number of all samples of all classes in the region $[X_1, X_1 + x]$, and, n the total number of samples. The equations for $q_m(x)$ and $q(x)$ are similarly derived. An illustration of the threshold determination over an interval range for a for a data variable can be found in [9].

The value of x can be any continuous value with any amount of increment. But for calculation purposes, it would be better to have a certain discrete increment, which would add to make the next value. The discrete value of x can be calculated by dividing the difference between X_2 and $X1$ by the number of segments (the number of increment). We can add the quotient to the first number and subsequently add the quotient to the following numbers to generate all candidate values of x. For example, if $X_1 = 20$ and $X_2 = 85$, and we choose the number of segments equal to 5, then the quotient of the division of the two numbers (65 and 5), is 13. Then x contains $\{13, 26, 39, 52, 65, 78\}$ as an array of candidate thresholds.

B. Decision Rule Extraction

A rule for binary classification for threat determination is actually a set of rules or rules patterns, forming a decision tree. With social status data variables, there are possibly multiple steps (or branches or stumps) resulting in a rule tree, by which the best threat classification can be made. Considering numerous data variables for two threat classes, while accommodating multiple steps (branches) and selecting the best single data variable at each step, the conditional entropy for the i^{th} variable is defined in the equation 8.

$$S_i = p_i(0)S_{i0} + p_i(1)S_{i1} \tag{8}$$

Now, conditional entropy equations are defined for sample values of the variable, '0' and '1', respectively. Equation 9 indicates the conditional entropy for '0' sample values and equation 10 the conditional entropy for '1' sample values for the variable.

$$S_{i0} = -\{p_i(T|0) \ln[p_i(T|0)] + p_i(NT|0) \ln[p_i(NT|0)]\} \tag{9}$$

$$S_{i1} = -\{p_i(T|1)\ln[p_i(T|1)] + p_i(NT|1)\ln[p_i(NT|1)]\} \tag{10}$$

In equation 8, $p_i(0)$ is the ratio of the number of samples of '0' in the i^{th} variable and the total number of samples, which is N for this case, and $p_i(1)$ is the ratio of the number of samples of '1' and N.

For two classes with binary sample values, We can define a set of four unbiased conditional probabilities as follows:

- $p_i(T|0)$: a probability of occurrence of outcome T (Threat) under the condition that the sample value is 0,
- $p_i(NT|0)$: a probability of occurrence of outcome NT (No-Threat) under the condition that the value is 0,
- $p_i(T|1)$: a probability of occurrence of outcome T (Threat) under the condition that the sample value is 1 and,
- $p_i(NT|1)$: a probability of occurrence of outcome NT (No-Threat) under the condition that the sample value is 1.

From the conditional entropies for all the variables, the best variable would be of the lowest conditional entropy. From the entropy equation, the entropy is the highest when the probability is 0.5; therefore, a variable whose probability p is as far away as possible from the middle, 0.5, would assure the lowest entropy S, and thus is selected for the rule pattern (or branch or stump) for the decision rule. Hence, below we define the variable index for the i^{th} variable in equation 11 using the previously defined conditional probabilities.

$$I_i = |p_i(1|T) + p_i(0|NT) - 1.0| \tag{11}$$

An interpretation of the variable index can be made as follows: the closer is the index to 1.0, the better is the variable in making out a rule pattern. Therefore, a certain variable that makes the highest index value is selected as the best variable in rule pattern (or branch or stump) generation among the variables. The rule at the step is made from the highest value in the four conditional probabilities set.

Algorithm 1: Recursive Rules Patterns Extraction's Algorithm

Input: A binary 2-D Array (table)
Output: A set of rules pattern

```
Function Rules(Binary Data Set):
    if BinaryDataSet lenght > 0 then
        Compute Index Attribute
        Select Best Attribute
        Compute Four Conditional Unbiased
          Probabilities
        Derive and Store a New Rule Pattern
        Remove samples from Binary Data Set base on
          last rule pattern
        BinaryDataSet ← New Reduced Binary Data
          Set
        Rules(BinaryDataSet)
    else
        return Rules Pattern
    end
```

C. Rule Pattern and Margin Error

After the order of the variables by the variable index values is obtained, the best variable is used to generate a rule at the first step (or branch) using the highest conditional probability. Suppose the rule's probability of correct classification (which will be discussed below) is acceptably high. In that case, the samples in the data whose variable values meet the rule may be removed from the data from which the next step (branch) rule generation starts using the second-best variable. Namely, the first variable is not considered anymore, and the step of the rule becomes a leaf or stump in the decision tree structure.

If, however, the correct classification probability is not acceptable, there is no removal of any sample but, in the second step using the second best variable, the conditional entropy equations consider the conditions of the first and the second variables together. Therefore, while at the first step with the best variable alone, there are aforementioned four terms of probabilities: $p_1(T|0)$, $p_1(T|1)$, $p_1(NT|0)$, and $p_1(NT|1)$, but when classification by the first variable alone is not sufficient, there would be the following eight terms in the conditional entropy equation with the first and second best variables: $p_{12}(T|00)$, $p_{12}(T|01)$, $p_{12}(NT|00)$, $p_{12}(NT|01)$, $p_{12}(T|10)$, $p_{12}(T|11)$, $p_{12}(NT|10)$, and $p_{12}(NT|11)$.

If in the second step using the top 2 variables, with the above eight conditional probabilities, the classification probability is sufficiently high enough, the samples which meet the highest conditional probability would not be considered in the third step of rule generation using the third-best variable. However, the classification certainty is not high enough; the number of terms in the conditional probabilities for three variables (first, second, and third best) would be 16. This iterative process of variable selection and rule pattern generation continues until there is no sample left in the data. Combining the rule patterns at each step forms the decision-rule tree for the given data.

As mentioned above, at each step, a rule pattern can be generated using a variable, and we now derive the rule's probability or certainty in each step of rule pattern formation using the principle of maximum entropy.

The unbiased probability derived by the maximum entropy probability is defined in equation 12, where C is the conditional probability parameter. For example, if $C = T|0$, then the value of x is determined by the number of samples whose value of the variable is '0' among the T (Threat) outcome samples and, n is the those in both T (Threat) and NT (No-Threat) outcome samples.

$$< p_i(C) > = \frac{x+1}{n+2} \tag{12}$$

We can also compute the margin of error defined in equation 13 where p is the unbiased probability in equation12, and z is the z-score value for a percentage of the confidence interval.

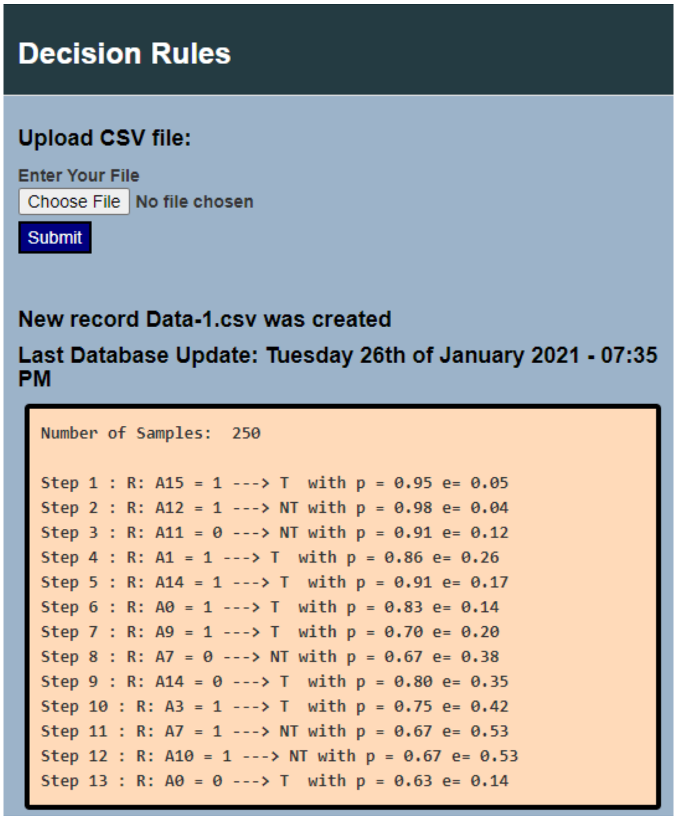

Figure 2. Display of rules pattern with 250 samples

From the z-score standard normal probability values, the value of z equals 1.65 for 90%, 1.96 for 95%, or 2.57 for 99%.

$$e_i(C) = \sqrt{\frac{p\,(1-p)}{n+2}} \cdot z \tag{13}$$

The algorithm 1 summarizes the extraction of rules patterns from a binary data set.

V. Example Data Analysis on Web-based Platform

In the web-based platform, we have created a web-interface for collaborators to share data and request the decision rule from the combined data. The interface web-page allows data upload, and receives the responds to rule generation requests from the collaborators.

In data collection from various sources, the minimum requirement for a data set to be used in data collection is that each contains at least one data variable. These variables can be, but are not limited to, grievance level, economic instability, the price of the main staple, change of leadership, number of civilian confrontations with the force, infrastructure instability, or inter-tribal conflict, and rivalry change [12]. The data variables can be of any data type such as text, number, and symbol; however, they will be eventually converted to numerical values via, for example, fuzzy logic and membership functions, especially for descriptive variables [13].

A. Polity IV Data Set

The decision assistance evaluation is conducted using the Polity IV data set available online through the Center for Systematic Peace (CSP) [14]. This center is devoted to innovating and developing research on political violence approaches in the dynamic global system's structural context. The center also provides support for scientific research and quantitative analysis in human relationships and socio-systemic development processes. Continuously, this center monitors political behavior in each of the world's major states, considering all those with a population of 500,000 (167 in 2014), reporting issues related to political violence and state failure. We are using data from the CSP since its researchers' record and report their findings openly and freely [14].

The Polity IV resource was prepared with open source information as a service to the research community by researchers in the Integrated Network for Societal Conflict Research (INSCR). The INSCR data resources are continuously verified with different data resources to ensure it is accurate, reliable and complete [14]. Polity IV project is a valuable resource for research. It has become the most widely used resource for comparative and quantitative analysis, especially for monitoring regime change and authority and the study of its effects.

The unit of analysis called polity can be defined as a political or governmental organization, society, or institution with an organized government, state, political body; or a simple definition that may be interpreted as subsets of authority class patterns. Polity IV combines essential information to be compatible with the state's fault investigation approach. Researchers have made a concerted effort to distinguish the characteristics of the regime and the authority of effective state policy from the use of organized and anti-regime armed force to challenge and, possibly, reject that authority.

B. Validation Process

To validate the web-based collaborative platform, we assume that two collaborators have sample data in commas separate value (CSV) format files from the Polity IV data set which is with 17 variables. One collaborator has a file of 250 samples and the other of 248. The implemented algorithm for rule generation can skip possible missing or corrupted sample data expressed as a 'NaN' value. Also, for illustration purposes, at each step of rule pattern formation, the implemented algorithm does not check if the rule pattern's probability is above a certain value; it assumes that each rule pattern at each step is good enough. Therefore, we expect to see at each step the rule is made from just one variable making that as a leaf node in the decision tree.

After the first collaborator submits the first data file, 'Data-1', as shown in the figure 2, the web-page displays in the Decision Rules section a message indicating that a new record has been successfully created, which means that the data is stored in the database of the web-server. It also displays the last day that the database is updated.

```
Number of Samples:  498

Step 1 : R: A15 = 1 ---> T  with p = 0.87 e= 0.05
Step 2 : R: A12 = 1 ---> NT with p = 0.95 e= 0.04
Step 3 : R: A11 = 0 ---> NT with p = 0.88 e= 0.13
Step 4 : R: A14 = 1 ---> T  with p = 0.95 e= 0.10
Step 5 : R: A1 = 1 ---> T  with p = 0.80 e= 0.35
Step 6 : R: A0 = 1 ---> T  with p = 0.86 e= 0.09
Step 7 : R: A7 = 1 ---> NT with p = 0.71 e= 0.34
Step 8 : R: A9 = 1 ---> T  with p = 0.71 e= 0.19
Step 9 : R: A8 = 1 ---> NT with p = 0.67 e= 0.53
Step 10 : R: A6 = 0 ---> T  with p = 0.57 e= 0.37
Step 11 : R: A16 = 1 ---> T  with p = 0.67 e= 0.53
Step 12 : R: A3 = 1 ---> NT with p = 0.67 e= 0.53
Step 13 : R: A0 = 0 ---> T  with p = 0.65 e= 0.09
```

Figure 3. Display of rules pattern with 498 samples

When a file is submitted, the database is updated, and the server requests the rule algorithm to be executed using the updated database. Figure 2 displays the output of the algorithm. The rule shows its 13 steps (branches or stumps). Considering that the database was previously empty, the number of samples in the database is equal to the number of samples in the data file 'Data-1', which is 250 samples.

The rule output is as explained above made just one variable at each step thus formed a leaf node. In descriptive rule interpretation, the first three rules on Figure 2 can be stated as follows:

- **IF** data variable #15 is above the threshold; the situation seems to be Threat with 95% certainty.
- **ELSE IF** data variable #12 is above the threshold, then the situation seems to be No-Threat with 98% certainty.
- **ELSE IF** data variable #11 is below the threshold, then there seems to be No-Threat with 91% certainty.

After submitting the second file, 'Data-2' from the assumed second collaborator, new rules are generated using the updated database. As shown in the figure 3, now the number of samples in the database is 498.

If we compare the rules from figures 2 and 3, we can appreciate that the rules extracted from the combined data (498 samples) have changed. In the first rule pattern, the probability of classification decreases from 95% to 87%. We can interpret that the collaborated data shares "its experience" in its local area. The classification and prediction are closer to the real-world situation with more diverse populace data from different locales.

VI. Conclusion

For expanded utilization of the information entropy-based rule generation algorithm, a web-based collaboration platform was developed and tested to collect data from collaborators and generate decision rules from a global perspective. Also, we discussed the algorithm itself with its essential components: variable binarization and the best variable selection using the entropy minimum principle, rule pattern formation for each step or branch using conditional entropy equations, and calculation of the accuracy or certainty of the rule itself at each step using the maximum entropy principle. We tested the web-server using the Polity IV data in a user interface for data upload, database update, and rules update. The web-based collaboration platform integrates a decision-assist system with multiple data sources so that learning and prediction of social unrest may be made in a global context by collaborated data collection and the updated decision rules.

References

[1] British Army. British army field manual, countering insurgency. *London: Ministry of Defense*, 1, 2009.

[2] Charles J Kim. Identification of symptom parameters for failure anticipation by timed-event trend analysis [power distribution systems]. *IEEE power engineering review*, 20(9):48–49, 2000.

[3] Weiming Shen, Qi Hao, and Weidong Li. Computer supported collaborative design: Retrospective and perspective. *Computers in Industry*, 59(9):855–862, 2008.

[4] Léon Bottou. From machine learning to machine reasoning. *Machine learning*, 94(2):133–149, 2014.

[5] Sriram Vajapeyam. Understanding shannon's entropy metric for information. *arXiv preprint arXiv:1405.2061*, 2014.

[6] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[7] ET Jaynes. Probability theory with applications in science and engineering. *Washington University*, pages 12–19, 1974.

[8] C.J. Kim and B. Don Russell. Classification of faults and switching events by inductive reasoning and expert system methodology. *IEEE Transactions on Power Deliv-ery*, 4(3):1631–1637, 1989.

[9] Charles Kim and Harald Schone. Machine reasoning for determining radiation sensitivity of semiconductor devices. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, pages 16–21, 2018.

[10] R. Christensen. *Entropy Minimax Sourcebook: Computer implementation*. Entropy Minimax Sourcebook. Entropy Limited, 1980.

[11] R. Christensen. *Entropy Minimax Sourcebook: Philosophical origins*. Entropy Minimax Sourcebook. Entropy Limited, 1980.

[12] Geoffrey Demarest. Property & peace: Insurgency, strategy and the statute of frauds. Technical report, Foreing Military Studies Office (ARMY) Fort Leavenworth KS, 2009.

[13] C. J. Kim. An algorithmic approach for fuzzy inference. *IEEE Transactions on Fuzzy Systems*, 5(4):585–598, 1997.

[14] POLITY IV. Polity iv: Regime authority characteristics and transitions datasets. Available online at: http://www.systemicpeace.org/inscrdata.html, 2017.