Aggregating Forecasts Using a Learned Bayesian Network

Suzanne Mahoney, Ethan Comstock, Bradley deBlois, and Steven Darcy
Innovative Decisions, Inc. 1945 Old Gallows Rd., Suite 207, Vienna, VA 22182
{smahoney, ecomstock, bdeblois, sdarcy}@innovativedecisions.com

Abstract

Under the Defense Advanced Research Projects Agency's (DARPA) Integrated Crisis Early Warning System (ICEWS), Innovative Decisions, Inc. (IDI) constructed a Bayesian network to combine forecasts produced by a set of social science models. We used Bayesian network structure learning with political science variables to produce meaningful priors. We employed a naïve Bayes structure to aggregate the forecasts. In both cases, IDI improved classification by intelligently discretizing continuous variables. The resulting network not only met performance criteria set by DARPA, but also out-performed each of the social science models across all types of forecasted events. We describe the construction of the aggregator as well as a set of experiments performed to explore the nature of the Bayesian EOI Aggregator's performance.

Introduction

For ICEWS, DARPA "seeks to develop a comprehensive, integrated, automated, generalizable, and validated system to monitor, assess, and forecast national, sub-national, and international crises in a way that supports decisions on how to allocate resources to mitigate them." (O'Brien, 2010) Lockheed Martin – Advanced Technology Laboratory integrated computational social science models to forecast country instability over a set of 29 countries. The forecasters used reports from open sources coded by event type along with political science variables to produce forecasts. IDI's Bayesian EOI Aggregator combined the forecasts into a set of predictions for five types of instability events of interest (EOIs): Ethnic-Religious Violence, Domestic Political Crisis, Insurgency, Rebellion and International Crisis.

We first describe the process used to construct IDI's Bayesian EOI Aggregator. We then present DARPA's performance criteria and summarize the Phase I ICEWS results. We next describe and present the results of several experiments conducted during Phase II. This is followed by a brief discussion.

IDI's Bayesian EOI Aggregator

The Bayesian EOI Aggregator as illustrated by Figure 1 is a Bayesian network (Pearl, 1988) with three types of random variables:

- Context variables from the political science literature
- EOIs being forecasted
- Forecaster estimates from social science models

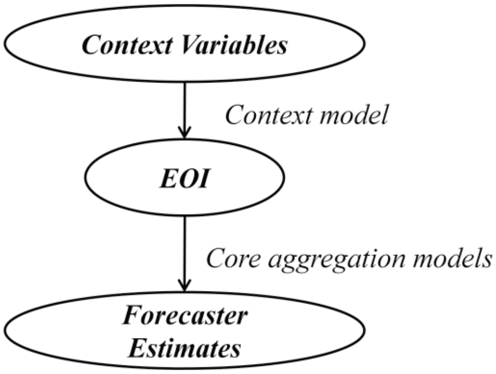


Figure 1: Abstract View of Bayesian EOI Aggregator

Constructing a Bayesian network has two foci: learning the structure and learning the parameters. (Heckerman, 1999) For ICEWS, data was available for calendar years 1998-2009 and 29 countries, so parameter learning was accomplished using Bayesian parameter learning (Russell and Norvig, 2003). Therefore, the focus of our work was to learn the structure of the network and its variables.

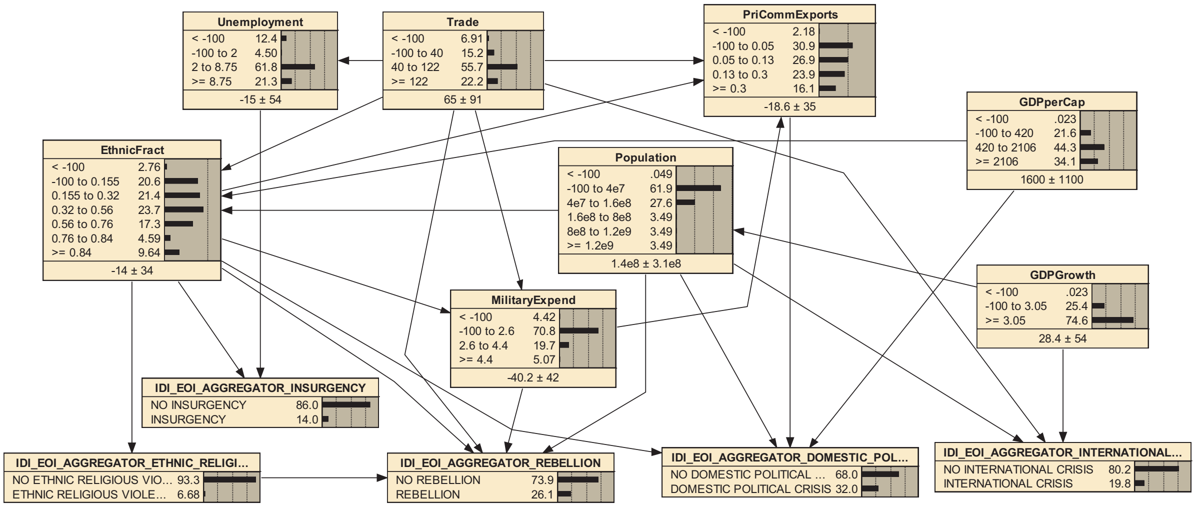


Figure 2 Context Variables with EOIs (Phase II Model)

Learning the Context Model

Figure 2 shows the Bayesian network for the context and EOI variables. The states of the context variables represent bins formed by discretizing the variables' continuous values. Because some data was systematically missing, we assigned it to the special state of -101. We obtained past data for the variables from public sources. Because the context model is predictive, we matched the context variables with EOI ground truth data provided by DARPA by lagging the context variables by at least one calendar month.

Learning the structure of the network presented two problems: 1) discretizing the context variables and 2) learning the structural dependencies among the variables. Given that interleaving discretization and model structure learning improves the performance of Bayesian network classifiers (Hoyt, 2008) we cycled through the following steps:

1) *Discretize the context variables*: By examining the training data, we subjectively clustered values for each of the eight continuous variables. Specifically, we discretized parents of a dependent variable to improve the discrimination among possible values of the dependent variable.
2) *Learn the Bayesian network structure*: We used Bayes Net Power Constructor (BNPC) to learn the structure of the context model. (Cheng, et al, 1998) Mutual information and conditional mutual information scores guide its construction process.
3) *Learn the parameters*: We used Netica^{TM} to learn the conditional probability tables.
4) *Test the resulting structure*: We examined the performance of a structure by exercising the resulting model to determine how well it predicted the EOIs using the training data.

The possible Bayesian network structures for a set of nodes is extremely large and structures generated are dependent upon the data set. Therefore, we used discretion in choosing an appropriate structure (Russell and Norvig, 2003). We chose a structure that performed well compared with other structures and met DARPA guidance: specifically DARPA encouraged us to avoid variables that could not be changed (e.g. mountainous).

Learning the Core Aggregation Model

Figure 3 shows the portion of the aggregator used to capture the EOI estimates for Ethnic Religious Violence. There are corresponding sets of nodes for the other EOIs.

The Core Aggregation Model is a set of conditional

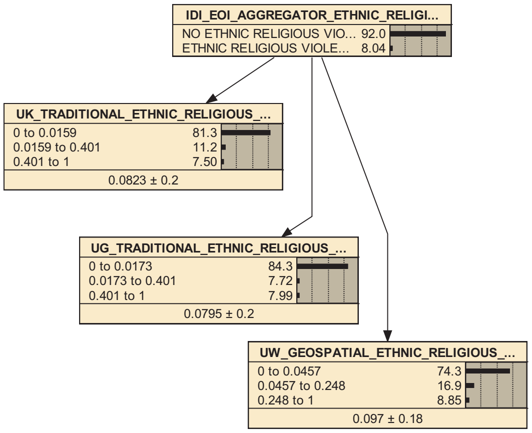

IDI_EOI_AGGREGATOR_ETHNIC_RELIGI...
NO ETHNIC RELIGIOUS VIO... 92.0
ETHNIC RELIGIOUS VIOLE... 8.04
UK_TRADITIONAL_ETHNIC_RELIGIOUS_...
0 to 0.0159 81.3
0.0159 to 0.401 11.2
0.401 to 1 7.50
0.0823 ± 0.2
UG_TRADITIONAL_ETHNIC_RELIGIOUS_...
0 to 0.0173 84.3
0.0173 to 0.401 7.72
0.401 to 1 7.99
0.0795 ± 0.2
UW_GEOSPATIAL_ETHNIC_RELIGIOUS_...
0 to 0.0457 74.3
0.0457 to 0.248 16.9
0.248 to 1 8.85
0.097 ± 0.18

Figure 3 Example of Core Aggregation Model

probability tables (CPTs), one for each forecaster's estimate for a given EOI. The nodes representing the forecaster estimates and their relationships with the EOI follow a naïve Bayes pattern (Duda and Hart, 1973). We selected this structure for several reasons: 1) In spite of its simplicity, naïve Bayes is known to perform well over a wide range of classification problems (Friedman, et al, 1997); 2) It is simple to implement and learn; 3) The structure made adding or removing a forecaster a straightforward exercise, independent of other forecasters.

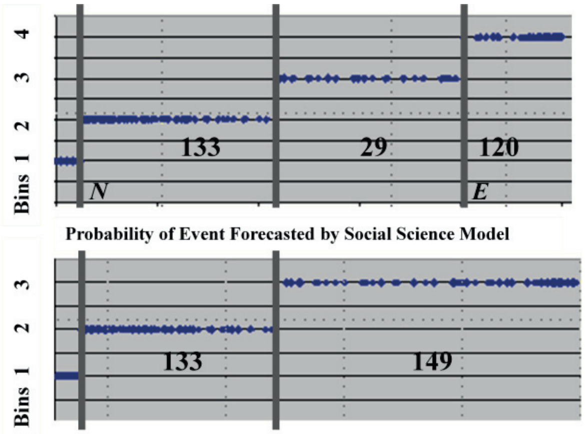
For each forecaster, training data consisted of forecaster estimates, continuous values ranging from zero to one, and ground truth EOI data for all 29 countries and 27 calendar quarters from 1998 through 2004. The challenge was to discretize each forecaster's estimates for an EOI to produce calibrated probabilities. Calibration entails adjusting forecasted probabilities so that the forecasted probability matches the probability of the event being forecasted.

Because the forecasters were constantly revising their models, IDI developed an application to automatically discretize the data. Early on we had achieved encouraging results using only two bins per forecaster estimate variable. These bins were based on the Receiver Operator Characteristic (ROC) curve of forecaster's estimates. Given that experience and knowing that naïve Bayes performance using continuous variables depends upon their discretization (Yang and Webb, 2002), we opted to create a finer discretization as follows:

1) On the assumption that a forecaster was at least partially calibrated, we used the training data to calculate: E, the average of the forecaster's estimates made in cases where the ground truth produced an EOI event; N, the average of the forecaster's estimates made in cases where the ground truth did not produce an EOI event; the average of E and N. These became the initial bin boundaries for the forecaster's estimates of an EOI.
2) If a bin had fewer than 30 data points, we combined it with a neighboring bin.
3) Next, we looked at the ratios of no EOIs to EOIs for predictions falling into the same bin. If the ratio did not decrease in going from a bin associated with lower probability of EOI estimates to a bin associated with higher probability estimates, we collapsed the bins into a single bin.

The process produced two to four bins for each forecaster providing estimates for each of the five given EOIs.

The example shown in Figure 4 graphically illustrates the initial and final bins for one of the forecaster models. In the two plots, the counts of predictions that fall within a bin are shown with the exception of the first bin whose count is 479. For each bin, the table presents the bin boundaries, normalized counts by ground truth, and ratios of no events to events. The normalized counts represent the CPT for the forecaster's estimates of the EOI. For a given prior, probabilities ranging from zero to one are reduced to three values, one for each bin. At the same time, all forecaster estimates above the low value of 0.422 favor EOI events over non-EOI events, thus reflecting the ground truth of those events. So, the CPT recalibrates the forecaster's estimates. This illustrative CPT is typical of the ones learned for the forecasters' estimates.



BIN	Upper Bin Boundary	Normalized Event Count	Normalized No Event Count	Event Ratio
Bin 1	0.049	0.10	0.78	7.99
Bin 2	0.422	0.07	0.20	2.78
Bin 3	1	0.83	0.02	0.026

Figure 4:Learning Discretization of Forecaster Estimates

To understand how well the resulting binning would perform, we ran a 10-fold validation test (Kohavi, 1995). K-fold validation involves randomly generating k test sets out of the training data, learning the parameters with the remaining data and then testing. The results provided an estimate of the model error. As shown in Table 2, the error rate can vary dramatically by EOI and among the 10 tests performed for the EOI.

Table 1: Error Rates for Selected EOIs

Test	Ethic Religious Crisis	International Crisis
1	0.01	0.12
2	0.02	0.08
3	0.00	0.12
4	0.01	0.06
5	0.00	0.15
6	0.00	0.20
7	0.02	0.09
8	0.03	0.15
9	0.04	0.12
10	0.05	0.08
Mean	**0.02**	**0.12**
Std Dev	**0.02**	**0.04**

Results for ICEWS Phase I

DARPA's performance measures include:

1) *Accuracy*: Proportion of predictions made that are correct. Note that this includes both no EOI events and EOI events. Part of the reason the accuracy is generally high for these EOI forecasters is that fewer than 20% of the quarters have EOI events. If one simply predicted no EOI all of the time, the accuracy would exceed 80%.
2) *Recall*: Proportion of EOI events that were correctly predicted. This measure applies only to EOI events. Note that one can easily get a 1.00 for this measure if one simply predicts an EOI event every time. The cost comes in accuracy and precision.
3) *Precision*: Proportion of correctly predicted EOIs among EOIs predicted. Like recall, this measure applies only to EOI events. Note that a high score in this measure may be obtained by only predicting EOI events that are almost certain. But, the cost is to lower recall.

DARPA's performance goals were 80% for accuracy and recall and 70% for precision. EOI events included Ethnic-Religious Violence (ERV), Domestic Political Crisis (DPC), Insurgency (Ins), Rebellion (Reb) and International Crisis (IC).

The ICEWS Phase I training and test data covered calendar quarters from 1998 to 2004 and 2005 to 2006 respectively. Although, the tests had high accuracy for all five EOIs, recall and precision were problematic in all cases. Table 2 summarizes the Phase I training and test results. Because accuracy is so easily achieved we present a single measure, the product of recall and precision, for comparing the performance of the forecasters with IDI's Bayesian EOI Aggregator. Given the DARPA performance goals, acceptable scores are greater than 0.56. The forecasters included Philip Schrodt of the University of Kansas (UK), Stephen Shellman of Strategic Analysis Enterprises (SAE) and Michael Ward of the University of Washington (UW).

In the training results, the Bayesian EOI Aggregator outperformed all the other models for every EOI. With the exceptions of IC and DPC, the Bayesian EOI Aggregator performed acceptably well on the test data. With the exception of DPC, the Bayesian EOI Aggregator outperformed other models for each EOI on the test data. When multiple models are making forecasts in the same direction, the Bayesian EOI Aggregator tends to make more extreme predictions that are closer to one or zero. Therefore it is not surprising that for DPC, an EOI for which all of the models performed poorly, the Bayesian EOI Aggregator performed even more poorly.

Table 2: Product of Precision and Recall for Phase I Data

	EOI	Logit Event UK	Logit Event SAE	Bayesian Event SAE	Spatial Networks UW	EOI Aggre gator
Training	Reb	0.75	0.78	0.55	0.57	0.88
	IC	0.37			0.18	0.62
	Ins	0.38	0.72	0.13	0.63	0.92
	ERV	0.79	0.75	0.50	0.25	0.81
	DPC	0.58	0.44	0.20	0.29	0.68
Test	Reb	0.39	0.80	0.53	0.40	0.83
	IC	0.25			0.15	0.37
	Ins	0.10	0.47	0.36	0.17	0.60
	ERV	0.48	0.50	0.27	0.31	0.64
	DPC	0.08	0.19	0.16	0.06	0.07

Phase II Experiments

In an effort to better understand and improve IDI's Bayesian EOI Aggregator, we posed a number of questions.

1) In Phase I, we used only four bins to successfully aggregate the forecasters' estimates. Does increasing the number of bins improve performance?
2) How does the performance of the Phase I aggregator compare with other approaches to aggregation such as simple averaging?
3) In Phase I, we used a Context Model to provide priors for the Bayesian EOI Aggregator. How critical are those priors to the results?

To investigate these questions, IDI used Phase II data. Unlike Phase I, DARPA required monthly predictions in Phase II for all 29 countries. The Bayesian EOI Aggregator was trained using data for the years 1998 – 2007 and tested with the years 2008 – 2009. We used early versions of the forecaster models so the results we present do not reflect their current performance.

Discretization Approaches

To answer the first experimental question, IDI implemented four discretization approaches to establish the bin boundaries for the forecasters' estimates. Bayesian parameter learning was then used to provide parameters. To respond to the second question, IDI implemented three approaches for combining estimates that assume that the forecasters' estimates are already calibrated and therefore do not need Bayesian learning to calibrate them. Table 3 lists the different discretization and aggregation methods with a description of each.

As before, tests showed high accuracy for all forecasters and discretization/aggregation approaches. Again, performance comes down to recall and precision. Table 4 presents their product for all EOIs and discretization/aggregation methods. In addition, the table also presents results for each of the individual forecasters: Michael Ward of Duke, Ross Schaap of Eurasia Group (EG), Philip Schrodt of Penn State (PS), and Stephen Shellman of SAE.

Table 3 – Discretization/Aggregation Methods

Method	Description
ROC Curves	A single bin boundary is set by taking the point in the ROC curve closest to the upper left hand corner of the graph.
Averages – 4 Bins	This is the default discretization method described earlier.
Averages – 6 Bins	This approach expands the Averages-4 approach by adding additional bin boundaries halfway between 0 (or 1) and the next bin boundary.
WPKID	WPKID (Yang and Webb, 2002) balances the number of bins and cases per bin. The number of bin boundaries is roughly the square root of the number of cases.
No Calibration – 3 Bins	This approach assumes the forecaster is calibrated. With three bins, the probabilities produced by a forecaster take on values of .167, .500, or .833.
No Calibration – 9 Bins	Similar to No Calibration – 3 Bins but more granular.
Average	Forecasts are simply averaged .

In considering performance across the EOIs, the calibrating approaches perform better than any single forecaster. Second, a poor forecaster makes little difference for calibrating approaches. Third, calibrating approaches outperform non-calibrating ones.

Having many bins does not necessarily produce better performance. In particular, WPKID was not found to perform better than the other calibration methods. This is due to the fact that the sets of probabilities being discretized and calibrated are generally clustered near zero and one. As a result the WPKID bins near zero and one have similar event to non-event ratios, and the benefit of having many bins is lost. This explains why all of the calibrating methods perform similarly.

How Various Context Models Impact Performance

The primary function of the context model within the Bayesian EOI Aggregator is to set the prior probability for each EOI. To better understand the contribution of the context model to the performance of the Bayesian EOI Aggregator, we constructed context models using different strategies and compared the results. Table 5 describes the context models.

Table 5: Description of Context Models

Model	Description
Uniform Priors	For each EOI the prior probability was set to 0.50.
EOI Base Rate Priors	For each EOI the prior probability was set to the base rate for that EOI across all countries.
Country Specific	This context model has just one variable: Country. The prior probabilities, learned from data, are specific to each country for each EOI.
Original Context	The context model structure was learned from data during Phase I. Its probability distributions were relearned from the monthly training data.

Table 6 shows the performance for the different approaches for developing priors. For all context models, the Averages – 4 method was used as the aggregation approach. As before, accuracy is not a discriminator, so we use the product of recall and precision in recognition of the trade-off between the two.

Country-specific context performs about the same as other approaches: This supports the belief that the context model may be simply learning to discriminate among

Table 4: Product of Recall and Precision for Discretization/Aggregation Methods and Forecasters

EOI	Calibrating Approaches				Non-Calibrating Approaches			Forecasters			
	ROC	Av4	Av6	WPKID	No Cal - 3	No Cal - 9	Average	Duke	EG	PS	SAE
Reb	0.77	0.74	0.75	0.73	0.74	0.75	0.78	0.81	0.58	0.32	0.76
IC	0.35	0.41	0.40	0.41	0.41	0.41	0.06	0.09		0.03	
Ins	0.30	0.26	0.22	0.27	0.03	0.05	0.03	0.06	0.00	0.06	0.05
ERV	0.60	0.58	0.57	0.60	0.02	0.42	0.04	0.78	0.00	0.13	0.25
DPC	0.30	0.22	0.22	0.29	0.08	0.14	0.08	0.07	0.00	0.03	0.37

Table 6 - Performance Comparison of Context Models Using Product of Recall and Precision

Event of Interest	Uniform Priors	EOI Base Rate	Country Specific	Original Context
Rebellion	0.74	0.75	0.74	0.74
International Crisis	0.15	0.10	0.51	0.41
Insurgency	0.28	0.24	0.24	0.26
Ethnic Religious Violence	0.63	0.61	0.63	0.58
Domestic Political Crisis	0.29	0.33	0.23	0.22

countries as well as the belief that one can use structural variables to model a country. Interestingly, with the exception of International Crisis, the uniform and EOI base rate priors' performances are comparable with those of the context model.

Discussion

Ensemble learning algorithms construct a set of hypotheses. On test cases, each hypothesis 'votes' for the classification (Dietterich, 2002). Stacking occurs when multiple base classifiers are combined using a learned meta-level classifier (Wolpert, 1992). The ICEWS modeling approach is an example of stacking with the Bayesian EOI Aggregator serving as the meta-level classifier.

Given that the probability distributions produced by any one forecaster tend to be bunched near its extremes with few forecasts in the middle, we have shown that any reasonable approach to discretizing the probability distributions of these forecasters works well. We have shown that by recalibrating the forecaster probabilities, the Bayesian EOI Aggregator generally out-performs any one other forecaster as well as non-calibrating approaches such as averages.

Although we have demonstrated that the Phase I Context Model could be readily replaced with a country-specific context model with minimal impact on performance, the performance of the context model indicates that a country may be effectively represented by a set of descriptive variables from the political science literature. It also demonstrates that an iterative discretization approach works well with political science variables.

Acknowledgements: This research was supported by AFRL under prime contract number FA8650-07-C-7749 Purchase Order: 4100054137. We thank Sean O'Brien and Dennis Buede as well as the anonymous reviewers for their helpful comments in reviewing the paper.

References

Cheng, J, D. Bell, W. Liu 1998. Learning Bayesian Networks from Data: An Efficient Approach Based on Information Theory. Published Online: http://www.cs.ualberta.ca/~jcheng/bnpc.htm

Dietterich, T. 2002. Ensemble Learning. In *The Handbook of Brain Theory and Neural Networks, 2nd Edition.* 405-408. Cambridge, MA: MIT Press.

Duda, R., and Hart, P. 1973. *Pattern classification and scene analysis*. Wiley and Sons, Inc.

Heckerman D. 1999. A Tutorial on Learning with Bayesian Networks. In *Learning in Graphical Models*, 305-354. Cambridge, MA: MIT Press.

Hoyt, P. 2008. Discretization and Learning of Bayesian Networks using Stochastic Search, with Application to Base Realignment and Closure (BRAC), Ph.D. diss., School of Information Technology, George Mason University, Fairfax, VA.

Friedman, N., Geiger D. and Goldszmidt M. 1997. Bayesian network classifiers. In *Machine Learning,* 29:131-163.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.* 1137–1143. San Francisco, CA: Morgan Kaufmann.

O'Brien, S. 2010. Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research. In *International Studies Review 12,* 87–104

Neopolitan, R., 2000. *Learning Bayesian Networks*, Upper Saddle River, NJ: Pearson Education.

Pearl, J., 1988. *Probabilistic Reasoning in Intelligent Systems*. San Francisco, CA. Morgan Kaufmann.

Russell, S., and Norvig, P., 2003. *Artificial Intelligence, A Modern Approach, Second Ed.* Upper Saddle River, NJ: Pearson Education.

Wolpert, D. 1992. Stacked generalization. *Neural Networks, 5,2*:241-259.

Yang, Y, G. Webb 2002. A Comparative Study of Discretization Methods for Naïve-Bayes Classifiers. In *Proceedings of PKAW 2002*, 159-173. Tokyo, Japan.