USING MACHINE LEARNING TO IMPROVE THE EFFICIENCY AND EFFECTIVENESS OF AUTOMATIC NUCLEAR EXPLOSION MONITORING SYSTEMS

Michael J. Procopio, Christopher J. Young, and Jennifer E. Lewis

Sandia National Laboratories

Sponsored by the National Nuclear Security Administration

Award No. DE-AC04-94AL85000/SL08-IRP-NDD02

ABSTRACT

An analysis is performed on the seismic-event data processed from 1999 through 2009 by the International Data Centre (IDC) of the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO). The analysis shows that the overall quality of the final IDC bulletin is excellent, but achieving this level of quality requires a significant amount of analyst effort. In particular, automatic processing produced 421,244 origin hypotheses, or about 118 per day over the nearly 11-year period of operation. Of these, 224,643 (53%) do not survive analyst review while the remaining 196,601 (47%) are approved. In addition, analysts build another 30,606 origins (13% of analyst approved total) that the automatic processing missed. Thus analyst-approved origins occurred at a rate of about 64 per day. It is evident that significant improvement of the automatic system, and hence decrease in the analyst workload, is possible both by decreasing the number of false events as well as by decreasing the number of missed real events.

Previous work in analyzing a much smaller portion of the IDC data (Gauthier, 2009) found that some attributes in the data appear well-suited for discriminating among the different families, or categories, of events. That evaluation used principled but ad-hoc methods to show a basic ability of certain attributes, such as number of stations and signal-to-noise ratio, to predict the true category of a particular detected origin, which may or may not be valid. In particular, a method was devised that identified with very high confidence many “false origins” which did not survive analyst review in the final Reviewed Event Bulletin (REB) table.

Building off the premise that there are predictive features in the origin and arrival data, the study herein proposes the use of modern machine learning and data mining methods to train models on previously labeled data found in existing archives. These models are then used to identify events produced by the automatic system which would otherwise be rejected after manual analyst review. In contrast to previous work in this area, the proposed methods are able to consider multiple data attributes simultaneously. Such an improved event detection system would reduce the analyst burden required in reviewing the events, the majority of which do not survive analyst review in current automatic systems.

Our study shows that contemporary supervised machine learning algorithms, including Support Vector Machines and Random Forests, are in the majority of cases able to correctly identify events that an analyst would reject after manual review. However, more work is needed in the areas of data attributes, data mining algorithms, and cost-sensitive learning in order to improve event categorization rates to acceptable levels, without appreciable gain in the false positive rate (FPR) of the categorizer.

OBJECTIVES

The objectives of this work are three-fold. First, we perform an analysis of a ten-year archive of IDC data in order to assess the percentage of the events produced by the automated event detection system that are not confirmed by analysts (hereafter referred to as “false events”). Manual analyst effort must be spent on screening out such events (false alarms). This analysis also characterizes the number of events missed altogether by the automatic system (missed detections).

Second, we propose the application of supervised machine learning (ML) to train predictive models on data from large, tagged event libraries. We show that these models are capable of identifying false events produced by the automatic system, which can be automatically screened out without manual analyst review. We give several examples of machine learning methods applied to this problem, ranging from simple methods yielding highly interpretable models, to more elaborate methods whose models may be more difficult to interpret.

Finally, based on the empirical results, we suggest a roadmap to help guide future research in this area. This work will improve the performance of machine learning methods on the task of screening out false events from the automatic detection system.

RESEARCH ACCOMPLISHED

Introduction

The verification regime of the CTBT includes the International Monitoring System (IMS) and the IDC. One purpose of this paper is to review the quality of the IDC bulletin of waveform-technology (seismic, hydroacoustic, and infrasonic) detected events throughout the history of operations to determine if there are characteristics of the data that could be utilized to understand and improve automatic seismic-event processing.

Previous work analyzing a month of IDC data (Gauthier, 2009) found that some attributes of the data appear well-suited to assist in discriminating among the different families, or categories[1], of events. That evaluation used principled but ad-hoc methods to show a basic ability of certain attributes, or features, to identify the true category of a particular automatically detected origin, which may or may not be valid. In particular, a method was devised that identified with very high confidence many “false events” which did not survive analyst review in the final REB table.

The study presented here uses modern machine learning and data mining methods to train models on labeled data drawn from the full IDC dataset (approximately 10 years), and then uses those models to make predictions on separate data whose categories are not known to the machine learning models. The results of the models’ predictions versus the known ground truth are compared and can be measured quantitatively.

IMS Network and IDC Processing

The current IMS network of seismic (primary and auxiliary), hydroacoustic, and infrasonic sensors provides good overall global coverage, given the limited distribution of land area (Figure 1). The IDC pipeline processes waveforms to produce detections, which are then associated with one another to create events. The pipeline is designed to meet requirements for both responsiveness and quality of the event bulletin. Three automatic event lists are produced with increasing quality as more data becomes available; these lists are known as Supplementary Event Lists and are referred to as SEL1, SEL2, and SEL3. The final of these (SEL3) is then reviewed by analysts to produce a high-quality REB. For this study, we focus on differences between SEL3 and REB.

[1] Although the terms “classifier” and “classification” are standard in machine learning, data mining, and related disciplines, we adopt the equivalent terms “categorizer” and “categorization” in this paper to avoid confusion with similar terms used in the nuclear explosion monitoring field.

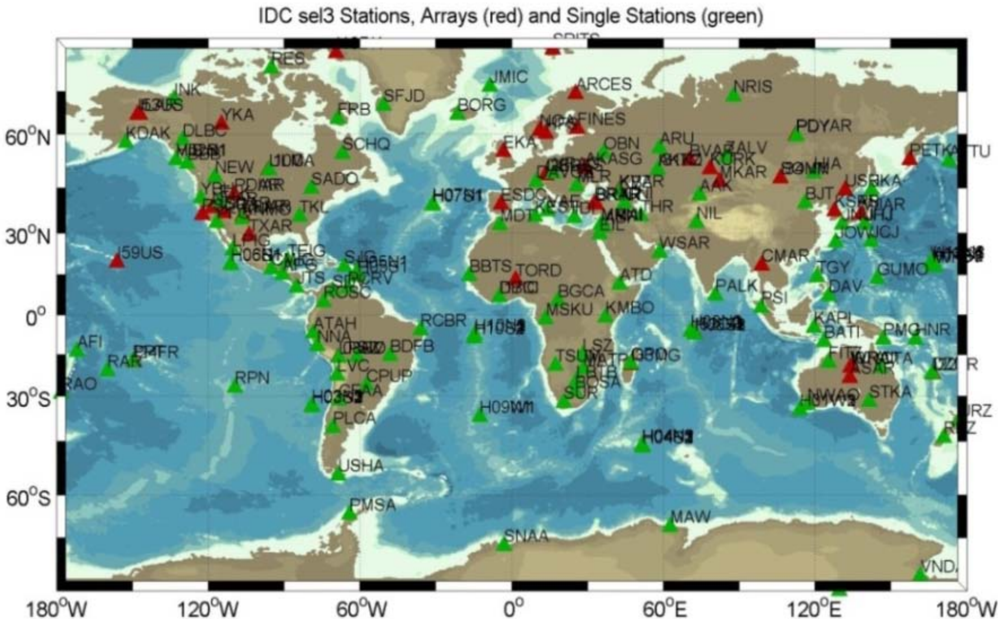


Figure 1. The IMS network, including seismic, hydroacoustic, and infrasonic stations

Analysis of IDC Event Catalogs

On average, only 47% of the automatically built events in SEL3 survive analyst review. This proportion seems to be fairly stable over the operating history of the IDC, despite the addition of new stations. Conversely, those events that are approved by analysts form approximately 87% of the REB. The additional 13% are newly built by the analysts, though they may make use of arrivals that were included in SEL3 events. Figure 2 shows the distribution of events into SEL3 and REB by year. This distribution is generally consistent from year to year (disregarding 1999 and 2009, for which only partial data was available).

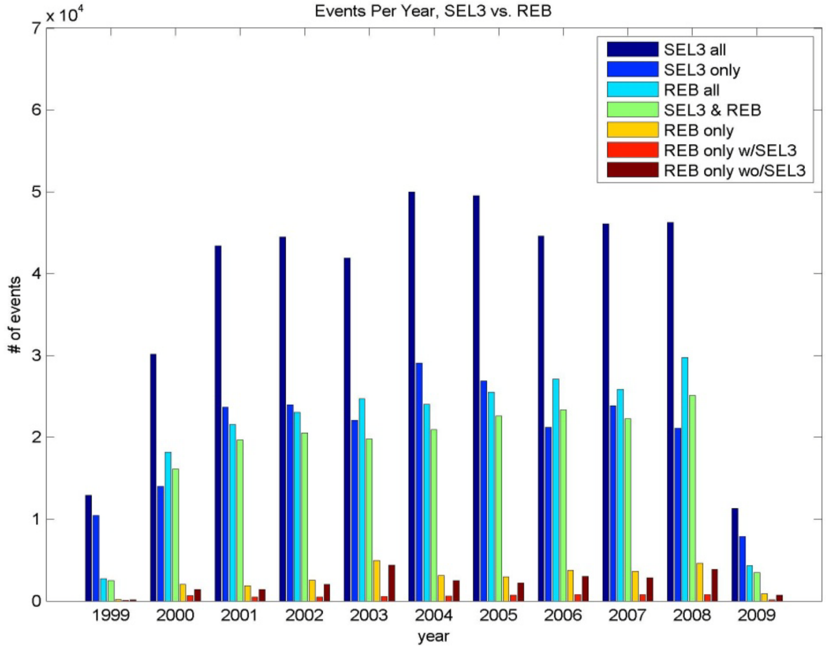


Figure 2. Categorization of automatic and analyst reviewed events for 1999-2009.

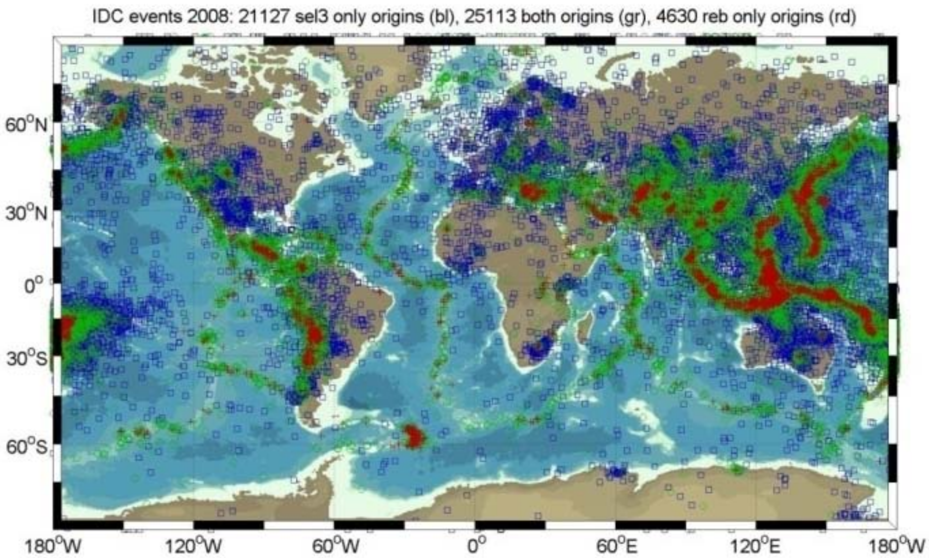

IDC events 2008: 21127 sel3 only origins (bl), 25113 both origins (gr), 4630 reb only origins (rd)
60°N
30°N
0°
30°S
60°S
180°W
120°W
60°W
0°
60°E
120°E
180°W

Figure 3. Comparison of automatic and analyst review events, 2008.

REB events occur almost entirely along the plate boundaries, as would be expected for earthquakes (Figure 3). SEL3 events that end up in the REB follow the plate boundaries, but not as reliably. Rejected SEL3 events can be very far from plate boundaries indicating false detections and/or false phase identifications.

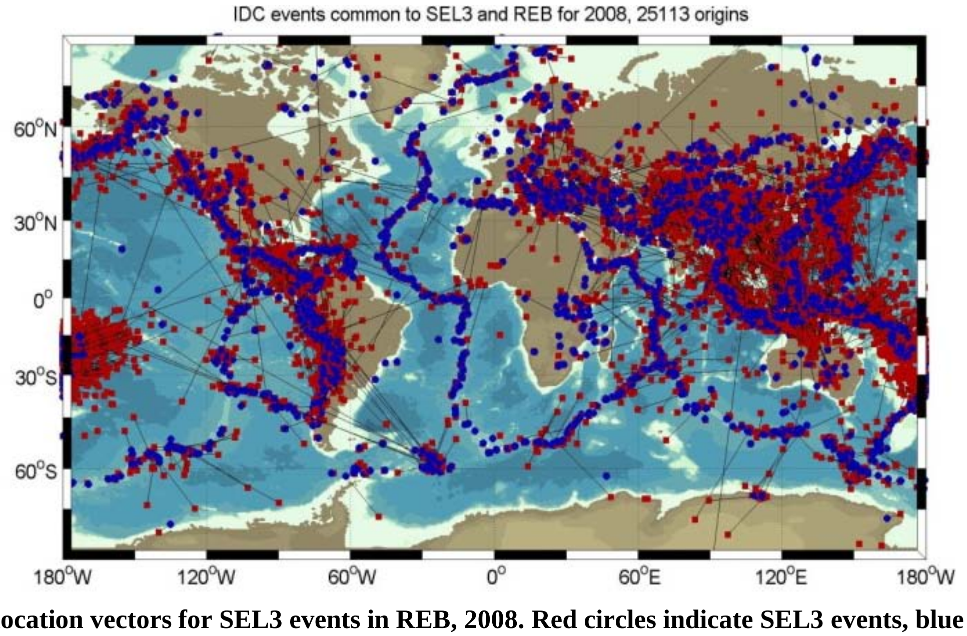


Figure 4. Relocation vectors for SEL3 events in REB, 2008. Red circles indicate SEL3 events, blue circles indicate REB events.

Many SEL3 events that are approved by analysts are shown to have large relocation vectors suggesting problems with probability of detection calculations in the Global Associator (GA) software (Figure 4). Predicted azimuth and slowness should show large variations between SEL3 and REB locations, with the smaller residuals corresponding to the REB-adjusted locations. GA grid points nearby the REB-locations would have been considered during the association step, but further ones were preferred. This problem may indicate the need for better azimuth and slowness calibration.

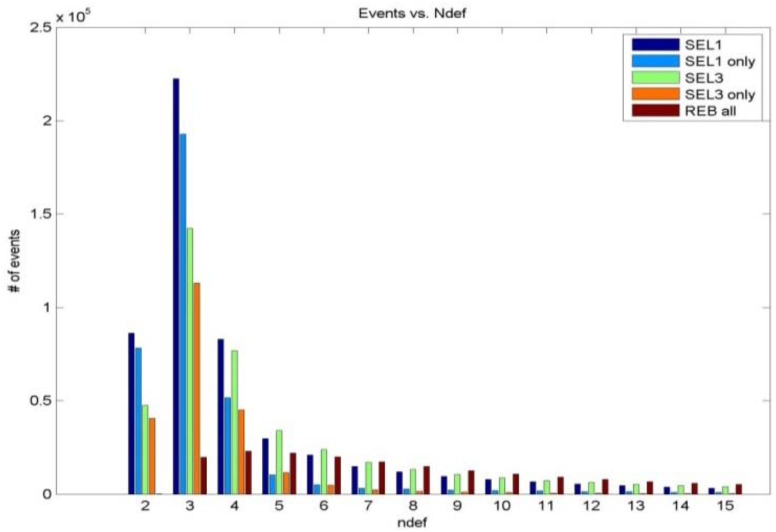


Figure 5. Number of events vs. number of location-defining phases, automatic vs. analyst reviewed, 1999-2009

The event definition criteria for automatic events is a combined weight of 3.55 (seismic travel time at a single station has a weight of 1.0), implying that an event can be built with time-defining phases from as few as two stations if azimuth and slowness are used. By comparison, analyst approved events must have a weight count of 4.6 or more and must have defining phases from 3 or more primary seismic stations (for terrestrial events). SEL1, with no auxiliary or late-arriving seismic data is dominated by 2 and 3 station events (Figure 5). However, 91% of 2 station and 87% of the 3 station SEL1 events do not result in REB events, so the payoff for building these events is small. SEL3 reduces those percentages somewhat to 85% and 79%, presumably due to the additional data that becomes available, but clearly a large amount of analyst effort is being used to screen these marginal events, suggesting that the automatic system may be building too many of them.

Of the 44,993,580 arrivals present in the IDC dataset considered in this study, only 4,049,843 (9%) are associated with SEL3 events. The total number of REB associated arrivals is 3,651,404, but this includes new arrivals added by the analysts. Thus, the vast majority of automatic detections are never used to produce REB catalog events. This disproportion is intentional because the consequences of missing an important detection are far more costly than those of producing a false detection. However, stations with anomalously high proportions of automatic arrivals to REB arrivals may be good candidates for further tuning.

Automatic Detection System Event Categories

The automatic detection system produces an event bulletin that gets filtered several times before finally resulting in the final list, SEL3 (referred to as the set SEL3-ALL in the results). The events in SEL3-ALL undergo analyst review, which divides the events produced by the automatic bulletin into two categories. The first category comprises all incorrectly categorized detections, known as False Automatic Origins (Gauthier, 2009). Because these origins were incorrectly predicted by the automatic system, they are labeled here as False Positives. These events go on to get screened out by the analyst; hence they do not survive analyst review (Figure 6).

The remaining events that do not get screened out are known as Good Automatic Origins (Gauthier, 2009), and are labeled as True Positives. Because these events endure through to the REB, they are labeled as SEL3&REB (i.e., those events in the intersection of the sets SEL3 and REB).

There also exists a set of events that should have been detected by the automatic system, but were not. These events are ones manually identified by an analyst and added to those surviving analyst review from the automatic system. These events, known as Extra Analyst-Built Origins, are labeled as False Negatives (or missed detections), because the automatic system incorrectly failed to identify them as true events.

The events surviving analyst review are combined with the additional events manually identified by an analyst; these are combined together to form the REB, and this set of events is referred to as REB-ALL. We note that both sets SEL3&REB and REB-ONLY can be further divided into subcategories (Gauthier, 2009); this is not represented here, but will be considered in future work.

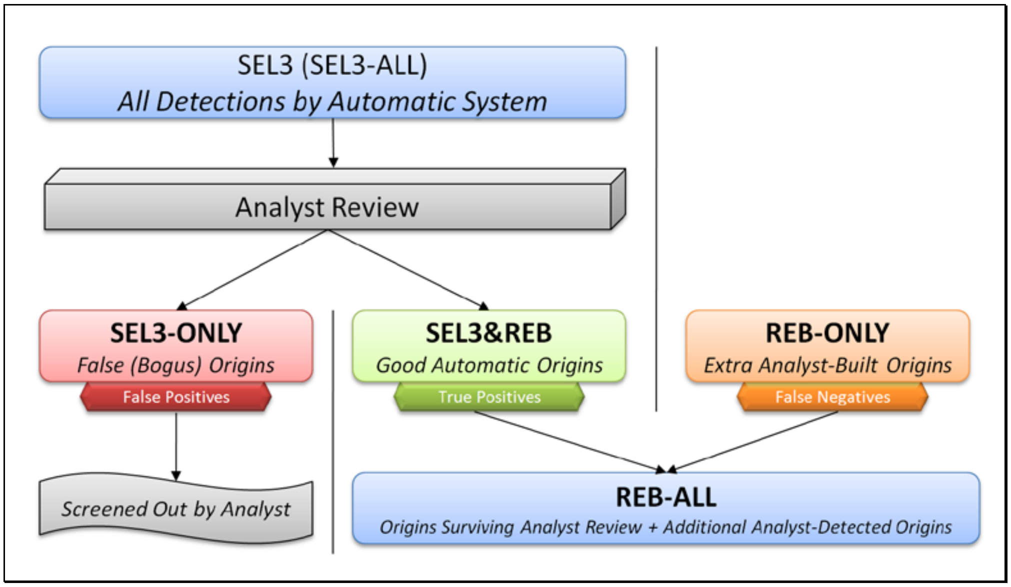


Figure 6. Overview of Automatic Detection System Event Families

Scenarios where valid measures exist for true positives, true negatives, false positives, but not true negatives are common in information retrieval. Traditionally, such scenarios are evaluated in terms of *precision* and *recall*, which measure the effectiveness of such a system. Type 1 and Type 2 error rate can also be used to characterize performance. In future work, the performance of the automatic system can be evaluated in these terms.

Use of ML Methods to Screen out False SEL3 Events

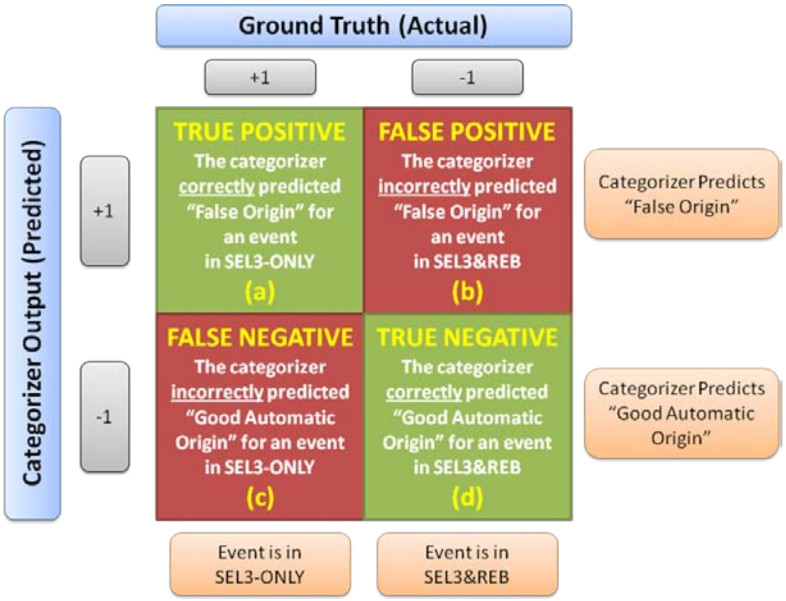
We propose the use of ML algorithms to process event features to try to reduce the number of false automatic events that IDC analysts must screen. Machine Learning methods can accommodate as many features as the researcher wants to evaluate in order to provide the best categorization performance.

For this initial empirical study for determining the feasibility of the machine learning approach on reducing analyst burden in the automatic detection system, we concentrate specifically on the problem of screening out events (False Origins) that an analyst would otherwise have to screen out. Thus, the problem we address is the automatic *discrimination,* or separation, of events which would eventually be assigned to the set SEL3-ONLY from those in SEL3&REB (Figure 6). Thus, the problem as framed is a two-category or *binary* problem, where the categorizer (i.e., the model learned on training data) predicts only one of two possible category labels for a given event in SEL3-ALL. The task of using ML methods to identify missed detections of the automatic system is not considered in the study presented here.

In our approach, we train supervised machine learning models using various algorithms on half of the tagged origins from SEL3-ALL, i.e., origins whose category is known. The model is then evaluated over the remaining test data set, i.e., the set of data disjoint from the training data set used to train the models initially. The goal is for the categorizer to correctly predict the true category (either SEL3-ONLY or SEL3&REB) for as many test origins in SEL3-ALL as possible. Because the ground truth category for each of these test origins is known, the performance of any particular categorization method can be evaluated by comparing the categorizer's predicted output with the known ground-truth labels.

Evaluation Approach: Contingency Table

We evaluate the trained ML models using the well-established contingency table approach, seeking to reduce the overall error rate (combination of false positive and false negative rates [FNRs]). Figure 7 shows the contingency table, also known as a *confusion matrix* in the machine learning and statistical literature, which characterizes all four possible outcomes of a binary prediction in a two-category problem.



		Ground Truth (Actual)		
		+1	-1	
Categorizer Output (Predicted)	+1	TRUE POSITIVE The categorizer correctly predicted "False Origin" for an event in SEL3-ONLY (a)	FALSE POSITIVE The categorizer incorrectly predicted "False Origin" for an event in SEL3&REB (b)	Categorizer Predicts "False Origin"
	-1	FALSE NEGATIVE The categorizer incorrectly predicted "Good Automatic Origin" for an event in SEL3-ONLY (c)	TRUE NEGATIVE The categorizer correctly predicted "Good Automatic Origin" for an event in SEL3&REB (d)	Categorizer Predicts "Good Automatic Origin"
		Event is in SEL3-ONLY	Event is in SEL3&REB	

Figure 7. Contingency Table showing the four possible outcomes in a binary (two-category) categorization problem, as applied to the SEL3 false event detection task.

Essentially, the categorizer can predict one of two categories, while the ground truth is also one of two categories, resulting in four possible scenarios. These scenarios form the cells in the matrix in Figure 7 and are referred to as true positives, false positives, false negatives, and true negatives. It is important to differentiate this evaluation approach and terminology from that in Figure 6. This is because *false positive* and *false negative* have different meanings in the context of categorizer prediction in the above SEL3 false event screening task (Figure 7), versus the meaning of *false alarms* and *missed detections* in the automatic event detection task (Figure 6).

Features Used

In general, ML approaches begin with identifying potentially useful data attributes, known as features. In this study, we used 18 features available from the IDC database tables for SEL3 events. These consist of nine origin-based features and nine arrival-based features, the latter which we group by event and then take the average to yield a single numerical value for each origin. Together these features, corresponding to a single observation (origin), form a *feature vector* for that origin. These features are summarized in Table 1. We note that selection of features is crucial for a high-performing machine learning categorization system, and further investigation of useful features is an important area of future work.

Table 1. Features considered in the study.

Origin Features				Arrival Features, Grouped by and Averaged by Origin		
#	**Name**	**Description**		**#**	**Name**	**Description**
1	ndef	*number of location defining phases*		**10**	avg(snr)	*signal-to-nose ratio*
2	nass	*number of associated phases*		**11**	avg(amp)	*amplitude, instrument corrected*
3	depth	*estimated depth*		**12**	avg(rect)	*rectilinearity*
4	sdobs	*standard error of observations*		**13**	avg(deltim)	*time uncertainty*
5	smajax	*semi-major axis of error*		**14**	avg(abs(timeres))	*time residual*
6	sminax	*semi-minor axis of error*		**15**	avg(delaz)	*azimuth uncertainty*
7	sdepth	*depth error*		**16**	avg(abs(azres))	*azimuth residual*
8	stime	*origin time error*		**17**	avg(delslo)	*slowness uncertainty*
9	numsta	*number of observing stations*		**18**	avg(abs(slores))	*slowness residual*

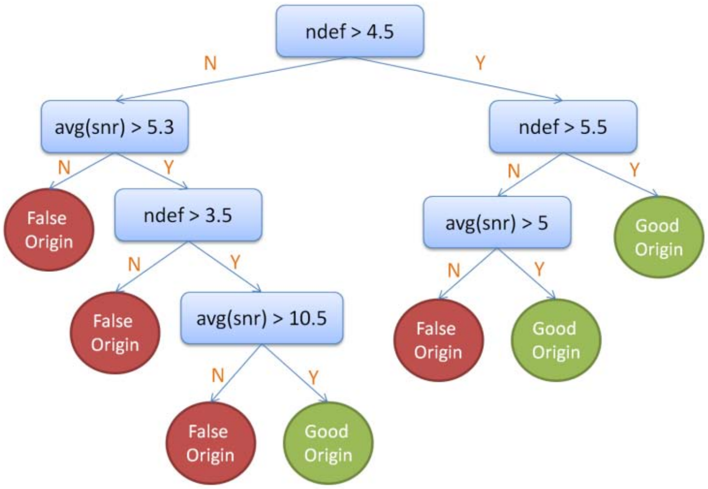


Figure 8. Example of an interpretable decision tree model providing non-linear two-category categorization for two input attributes.

Machine Learning Algorithms

In this study, we consider three machine learning algorithms, comparing the results to a first-order baseline method. Results for both of these models, as well as the more complex models below, are given in the experimental results shown in Table 2.

- **Baseline: Single Split on a Single Attribute**. For comparison, we developed a baseline method that makes categorizations based on splitting on single attributes; ndef and avg signal-to-noise ratio (SNR) are obvious choices for single attributes. The split location for the particular attribute is chosen simply as that which maximizes event categorization accuracy. For ndef, this yields the following model: ndef <= 4 is a false automatic origin, ndef >= 5 is a good automatic origin. For avg(SNR) the split logic is similar, but the threshold value is 5.5.

- **CART (Classification and Regression Trees) Decision Trees**. CART trees, due to Breiman et al. (1984), provide a simple mechanism of learning categorization rules from data. The resulting models, expressed as binary decision trees, can give non-linear categorization yet are generally interpretable (in contrast to the models yielded by other methods, such as Neural Networks). In this study, we learn a decision tree on two attributes, ndef and avg(SNR); the resulting tree is shown in Figure 8. In contrast to the single-split baseline, it allows for a more complex, non-linear separation of the two categories.

- **Random Forests**. The Random Forests technique, due to Breiman (2001), is a state-of-the-art machine learning ensemble method that combines *bagging* with *random trees*. This method is robust to noise, requires minimal parameterization and tuning, and performs well in the presence of possibly irrelevant features. Moreover, this method has a heavy theoretical basis. Here, this method is applied it to all 18 features simultaneously. We specified the number of trees, a parameter of the algorithm, as 30. The resulting model, composed of many simple decision trees, can be readily evaluated quantitatively but not easily visualized or interpreted.

- **Support Vector Machines (SVMs).** Like Random Forests, The SVM algorithm is a relatively recent method (Vapnik, 1995) and considered to be state-of-the-art. As opposed to decision trees, SVMs use kernel methods to achieve linear separation of data in a transformed feature space that may be of higher dimension than the original input data. The resulting decision boundary, when projected in the original space, can yield non-linear categorization, depending on the type of kernel used. In this study, we use the Radial Basis Function kernel, and chose parameters c = 2 and g = 2 based on 10-fold cross validation.

Experimental Results

A summary of quantitative experimental results from the study is given in Table 2.

Table 2. Experimental results summary.

Algorithm	Feature(s) Used	False Negative Rate	False Positive Rate	Accuracy
Baseline (Single-Split)	avg(SNR)	59.5%	9.0%	65.7%
Baseline (Single-Split)	ndef	11.6%	34.5%	76.9%
Decision Tree	ndef & avg(SNR)	13.8%	29.8%	78.2%
Support Vector Machine	All features	14.4%	27.8%	78.9%
Random Forest	All features	15.8%	18.2%	83.1%

We report the False Negative Rate, False Positive Rate, and Accuracy for each approach. The False Negative Rate and False Positive Rate metrics give a useful characterization of the breakdown of the different types of errors, while accuracy is a single-value summary statistic that weights both types of errors equally. Equations for each metric are given below in Table 3; also given is the formula in terms of the quadrants of the contingency table in Figure 7.

Table 3. Performance metrics used in the study.

Metric	Standard Equation	Equation (Figure 7)
False Negative Rate	$\frac{Number\ of\ False\ Negatives}{Number\ of\ True\ Positives\ +\ Number\ of\ False\ Negatives)}$	$\frac{c}{a+c}$
False Positive Rate	$\frac{Number\ of\ False\ Positives}{Number\ of\ False\ Positives\ +\ Number\ of\ True\ Negatives}$	$\frac{b}{b+d}$
Accuracy	$\frac{Number\ of\ True\ Positives + Number\ of\ True\ Negatives}{Total\ Number\ of\ Observations}$	$\frac{a+d}{a+b+c+d}$

The experimental data given in Table 2 provides a number of insights. The best performance in the study was given by the Random Forests method using all 18 features as listed in Table 1. This performance exceeded that of the Support Vector Machine by a statistically significant amount (5.3% increase in accuracy), notable since many problem domains lend themselves to a particular category of machine learning algorithm.

In comparison to naively splitting on a single attribute, the approaches which consider multiple attributes yield greater categorization performance. However, the best performing method (Random Forests, accuracy of 83.1%) compared to the single-split baseline using ndef (accuracy of 76.9%) represents only an 8% increase. It is clear that ndef alone is a very predictive feature, which follows intuition. Meanwhile, the other features selected enable statistically significantly higher categorization rates, but not drastically so. More predictive features are needed.

For all approaches, the proportion of types of errors (false positives versus false negatives) was not equal. In particular, the False Positive Rate was usually higher than the False Negative Rate. This is an important observation, because for monitoring, different types of errors do not have equal cost. In particular, any miscategorization associated with potentially missing an event has much higher cost than an error which simply results in additional analyst burden (i.e., the time required to review the event manually). Considering the types of errors separately results in a complex interpretation of the performance scores reported in Table 2. Although Random Forests had the highest overall accuracy, splitting on a single attribute—avg(SNR)—resulted in the lowest False Positive Rate. This result must be considered with that of the *trivial categorizer*, that is, the categorizer that simply predicts all “Good Automatic Origins.” Such a categorizer would have a zero False Positive Rate, but very poor accuracy, and of course, no real impact in terms of screening out events and reducing analyst burden.

In general, the overall accuracy rates and in particular the false positive rates yielded by the methods in this study need to be improved in order to be seriously considered for inclusion in an automated event detection system. We provide a number of recommendations for future research that could lead to more accurate categorization of events.

CONCLUSIONS AND RECOMMENDATIONS

Applying Machine Learning to utilize a large number of readily available features to screen events that analysts will reject can lead to improved performance, over a simplistic categorizer, but the results are still not good enough to use because of the possibility of screening real events of interest. We believe that significantly better results can be achieved by use of new and enhanced features, and especially by the use of *cost-sensitive learning* to bias the categorizer towards lowering the False Positive Rate (FPR, see Table 3) at the expense of the False Negative Rate (FNR), reflecting actual monitoring system goals.

An important objective from this study was to inform future research directions in order to make the machine learning approach more effective. Future work will fall under three main areas: identification of additional features to help inform categorization; improved categorization algorithms; and the use of cost-sensitive learning to help improve results. These areas are discussed below.

Useful features are the most important part of any machine learning categorization task. Poor categorization performance will occur if features that do not support the discrimination task are used. There are a number of ways to improve the features for this problem. First, the power of existing features can be improved by applying special scalings and transforms. Second, new features can be computed as functions of existing features (e.g., distance between first and second detecting stations). Finally, entirely new features can be extracted from raw waveforms; this is an area of ongoing research (Meyer et al., 2009).

Future work will involve additional experimentation with various approaches and the categories of algorithms in order to improve performance. An important consideration is that once the best general type of categorizer is identified, improvements will often be incremental among best categorizers, for the same feature set. Presently, decision trees and related ensemble methods such as Random Forests perform best. Other categories of algorithms will also be investigated, including probabilistic categorizers (Naïve Bayes), neural networks, and K-nearest-neighbor approaches. We are cautious to not draw strong conclusions in regards to best algorithms for this task, as this will be very heavily feature-dependent.

In this domain, like many domains, the penalty for one type of error (incorrectly predicting a “False Origin”) is much greater than the other type of error (incorrectly predicting a “Good Automatic Origin”). Machine Learning has *cost-sensitive* methods for handling unequal costs. Different error costs can be explicitly accounted for when training models in order to bias the model towards one type of error. Unequal error costs can also be considered when evaluating the performance of a categorizer. For example, if specific costs are known, a weighted average of the errors can be made, where one type of error receives more weight in the final score. A more elaborate evaluation technique is Receiver Operating Characteristic (ROC) analysis, where all possible error costs are considered; the resulting Area Under the ROC Curve (AUC) summary statistic is a robust evaluation metric (Provost et al., 1998).

REFERENCES

Breiman, L.(2001). Random Forests. *Journal of Machine Learning Research*, 45: (1), 5–32.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984). *Classification and Regression Trees*. Chapman and Hall, Boca Raton, FL.

Gauthier, J. H. (2009). Preliminary Analysis of the International Data Centre Pipeline, Sandia National Laboratories Technical Report #2009-4223.

Meyer, F., K. M. Taylor, D. Kaslowsky, M. Procopio, and C. Young (2009). Evaluation of Empirical Mode Decomposition and Chirplet Transform for Regional Seismic Phase Detection and Identification [abstract]. *Seismo.l Res. Lett.* 80: (2), 347–348.

Provost, F., T. Fawcett, and R. Kohavi (1998). The Case against Accuracy Estimation for Comparing Induction Algorithms. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.