SAND2021-0992C

A Minimally Supervised Event Detection Method

Matthew Hoffman, Sam Bussell, and Nathanael Brown

Sandia National Laboratories, Complex Systems for National Security, P.O. Box 5800, Albuquerque, New Mexico, 87185, United States of America
{mjhoffm, sbussel, njbrown}@sandia.gov

Abstract. Solving classification problems with machine learning often entails laborious manual labeling of test data, requiring valuable time from a subject matter expert (SME). This process can be even more challenging when each sample is multidimensional. In the case of an anomaly detection system, a standard two-class problem, the dataset is likely imbalanced with few anomalous observations and many "normal" observations (e.g., credit card fraud detection). We propose a unique methodology that quickly identifies individual samples for SME tagging while automatically classifying commonly occurring samples as normal. In order to facilitate such a process, the relationships among the dimensions (or features) must be easily understood by both the SME and system architects such that tuning of the system can be readily achieved. The resulting system demonstrates how combining human knowledge with machine learning can create an interpretable classification system with robust performance.

Keywords: Human-systems Integration · Bayesian Networks · Rare Events · Supervised Classification · Data Fusion · Machine Learning

1 Introduction

Due to their infrequent nature [1], rare events are difficult to model and detect due to a low number of positive ("event") cases relative to the number of negative ("non-event") cases. Thus, detection is regarded as an imbalanced classification problem which attempts to detect events with high impact but low probability. Rare events detection has many applications such as network intrusion detection and credit fraud detection [2]. We are concerned with rare events of interest, a subset of rare events that must also meet some "importance" criteria. That is, we are focused on problems where all interesting events are rare but not all rare events are interesting.

We describe a method for human-in-the-loop automated filtering and classification for more efficient labeling of data that contains an abundance of uninteresting observations. Our approach consists of a three-step method: (1) a modified ensemble technique acting as a novelty filter which labels uninteresting data, (2) SME tagging of the remaining unlabeled data, and (3) classification of the further reduced unlabeled data using a Bayesian Network (BN). We are specifically interested in problems with many event detectors that output nonnegative values as our features, where zero means nothing happened and larger values indicate alerts of greater

interest or concern. The novelty filter prioritizes the most interesting observations for SME review by assuming that the presence of more alerts (non-zero features) and/or rare alerts in a single observation make it more interesting and provides interpretable reasons for the novelty score. Using a BN for the final classification has several benefits. As probabilistic graphical models of a set of variables (corresponding to our detector features) and their conditional dependencies, BNs are more natively interpretable than most other machine learning methods. BNs have also been shown to perform well in the area of rare event detection. Previous applications have included intrusion detection as demonstrated by Benferhat, et al. [3] who used naive Bayes and Tree Augmented Naive Bayes classifiers; Cheon, et al. [4] who used a BN in ozone level modeling to automatically alert a forecaster when abnormal signals (ozone levels) are detected; and Wong, et al. [5] who used BNs for detecting disease outbreaks. BNs have additional advantages over other machine learning techniques [6]: they provide a natural way to handle missing data, facilitate learning about causal relationships between variables, are robust to overfitting, and can deliver good prediction accuracy even with small sample sizes.

To demonstrate extensibility of our approach to a variety of domains and problem sets, we evaluated against a generalized synthetic data set that is not tied to a particular use case. Although our data set is binary, we anticipate that the method can be equally applied to ordinal, continuous or categorical (via binary encoding) data or mixes thereof, insofar as larger values are of greater interest. Additionally, our method provides several benefits and differentiators relative to general classification problems where there is no initial labeling of data:

- Reduces SME's data review burden and assists the SME with data labeling
- Using a BN for classification allows detecting multivariate patterns in the positively tagged ("interesting") cases, tolerates and infers missing data, and natively provides classification likelihood and model fit estimates
- Both the novelty filter and the BN are relatively interpretable and explainable

2 Method

The method consists of progressive data classification steps as depicted in Figure 1. Step 0 shows the training-test partition normally used for a classification problem. In step 1, a novelty filter as described in section 2.1 is used to automatically label most of the data as *uninteresting* (identified by the highlighted blue region around a negative sign ['-']). The labels in the test data are treated as *final* system classifications (i.e., the BN will not override them in step 3). In step 2, SME(s) review the remaining unfiltered samples in the training set and label them (*interesting* data is identified by a positive sign ['+']). Finally, a BN is built and trained based on *all* labels from previous steps and used to classify the remainder of the test set as highlighted in step 3.

Fig. 1. Method steps with updated labels/classifications highlighted in blue

2.1 Novelty Filtering

We propose a weighted voting model $s_j = \sum_{i=1}^{N} w_i a_{ij}$ where a_{ij} denotes the value of feature i in sample j, and $w_i = W_i (c_i)^{-k}$, where c_i is the sum over all samples for the i^{th} feature, W_i is an *a priori* feature weight, and (positive) k determines the relative importance of rarer values. This model is appropriate for directed nonnegative features where larger values are more interesting, such as binary alarms, ordinal alert levels or continuous meter readings. It assumes that features with more frequent positives are less informative about abnormal conditions, and samples with more unusual features are more abnormal. Using $k = 1$ encodes that more active features are proportionately less interesting, and $W_i = 1\ \forall i$ gives equal feature importance otherwise.

Samples with score s_j falling below a threshold value are automatically labeled as uninteresting. This enables SME review of a smaller data set and training the Bayesian network on the filter-labeled uninteresting examples in the test set. While choice of threshold is subjective, a low initial choice can filter out many observations in an unbalanced set and can be iteratively updated. In section 3.1 we discuss (by way of example) the process of informing choice of threshold with domain information.

Key benefits of the filter are its interpretability and adaptability: the reason for a sample's score is clear based on its contributors, and *a priori* feature weights can be adjusted if score rankings or feature contributions conflict with domain knowledge. Scores and contributions may also assist the SME in tagging the remaining data.

2.2 SME Tagging and Bayesian Network Classification

SME tagging is accomplished via reviewing the multivariate samples in the reduced training set and manually assigning a class (e.g., True, False, or Red, Yellow, Green).

Upon completion of SME tagging, a BN is built and trained from all previously labeled data (including those in the test set classified by the filter) and used to classify the remaining samples. We use a Tree-Augmented Naive (TAN) Bayesian network, a restricted BN class which combines the simplicity of Naive Bayes with the ability to express the dependence among attributes in a Bayesian Network. It embodies a good tradeoff between the quality of the approximation correlation among attributes and the computational complexity in the learning stage. TAN relaxes the naive Bayes attribute independence assumption by employing a tree structure, in which each attribute only depends on the class and one other attribute. A maximum weighted spanning tree that maximizes the likelihood of the training data is used to perform classification [7][8]. Figure 2 shows a BN representation of the type that TAN creates [7].

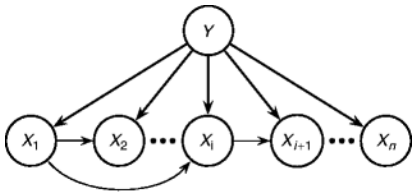

Fig. 2. TAN Bayesian Network

Key BN benefits for human-in-the-loop domain-informed classification include:

Native interpretability – Traditional machine learning (especially deep learning) still suffers from its black-box nature and many challenges remain to enhancing its interpretability and explainability [9]. One mitigation has been to use BNs to aid humans in interpreting the results of complex deep learning models [10]. As pointed out in [6] and [11], the visual nature of a BN as a directed acyclic graph can be used to communicate the underpinnings of a model via the causal relationships among the real-world features. For example, in figure 2 we can immediately see that the probability distribution of X_i is dependent on its parents X_1 and Y.

Explainability – BNs enable explainable classification via mutual information (describing how strongly features relate to the class variable), and provide metrics of classification confidence (e.g., probability of the assigned class) and per-sample model fit (e.g., log likelihood).

Tolerance for missing data – BNs natively tolerate and infer missing data features within a sample.

2.3 Iterative Workflow

Although we present and describe the novelty filtering and SME tagging processes sequentially, in practice (and especially on large data sets) it can be thought of as an iterative workflow with no defined entry or exit points as shown in Figure 3.

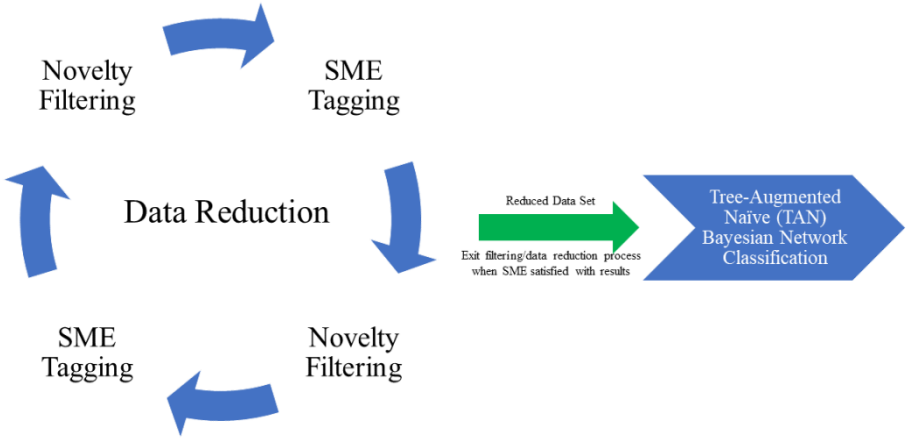

Fig. 3. Envisioned iterative method on larger data

The data reduction process prior to BN training and classification can begin and/or end with either novelty filtering or SME tagging, based on SME/analyst objectives and assessment of the results of each iteration.

3 Example Results

To provide a challenging example, we engineered a small data set with weakly correlated and sparsely positive binary features and highly unbalanced classes. We sampled 1000 observations with 10 features where the first 250 serve as the training set.

The first three features are independent random binaries with a positive rate of 20%, 10% and 5%, followed by seven ordered features each with a 10% chance of being positive if any of the prior three features in the same sample were positive and another 10% chance of being independently positive otherwise. Ground truth classes are synthetic and rule-based, where an observation belongs to the "interesting" class only if any of the following criteria are true:

1. Features 2 and 5 are both positive
2. At least two of the remaining first six features are both positive
3. At least two of the last four features are both positive

While this data generation scheme does not mimic any specific known data set, it provides the complexity we desired for our example in terms of feature sparsity, weak multivariate correlation, and unbalanced classes based on nontrivial multivariate patterns. The random draw used for this example had 13 "interesting" training set cases.

3.1 Choosing the Novelty Filter Threshold

Upon calculating novelty scores, one must determine the threshold below which observations are filtered out as uninteresting. Figure 4 depicts how the threshold score can be bounded by above and below.

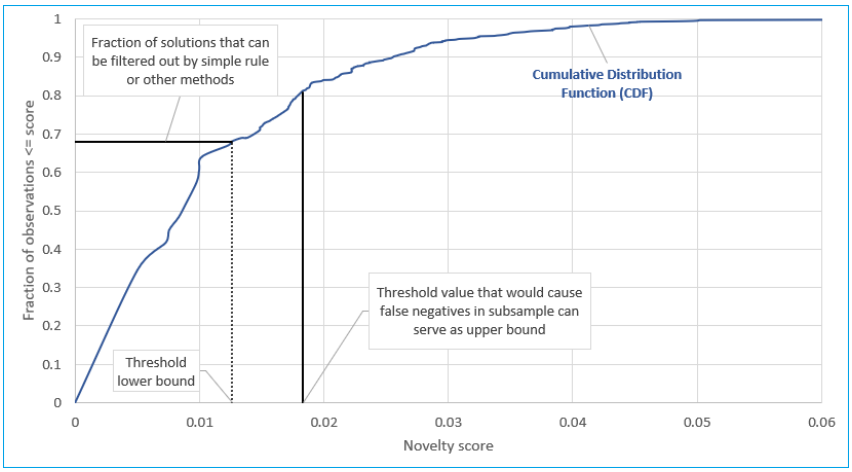

Fig. 4. Example of using domain information to bound novelty threshold.

Our lower bound in this example is found by examining the training set, which reveals that all interesting samples have at least two positive features. The upper bound is established by iteratively applying the novelty filter using various thresholds. On real data, domain-informed rules based on SME knowledge should be used and may be more sophisticated. While ultimately a matter of choice, picking a threshold value above the lower bound reduces the amount of data for SME tagging, and picking a value below the upper bound reduces the likelihood of false negatives. The threshold score chosen for this example was 0.017, resulting in labels of uninteresting for approximately 75% of the data and leaving 57 candidate samples in the training set for SME review. Note that we are using an empirical CDF based on the entire population, so the fraction of novelty scores less than X and P(x<X) are identical.

3.2 SME Tagging and TAN BN Classification

In practice, the next step would be SME classification of the remaining candidates in the larger training set (i.e., the 57 samples not filtered). For this illustrative example, SME training set classifications are simply assumed to match ground truth.

The TAN BN was learned from the fully labeled training set and the filter-labeled portion of the test set. The BN provides a probability for each classification. We define a "prediction level" as the minimum value of probability P(class = "interesting") required to tag a sample as interesting. Results of accepting the classifications as-is (i.e., using a 50% prediction level) are summarized in Tables 1 and 2.

Table 1. Confusion Matrix (50% prediction level)

		Predicted	
		0	1
Actual	0	707	11
	1	6	26

Table 2. Classification statistics at 50% prediction level

Accuracy	97.7%	**True Positive (TP) Rate**	81.3%
Precision	70.3%	**False Positive (FP) Rate**	1.5%
Recall	81.3%	**True Negative (TN) Rate**	98.5%
F1	75.4%	**False Negative (FN) Rate**	18.8%

In Table 3 we summarize the results by P(class = predicted class) given by the BN. These results show that most incorrect classifications occur at lower probabilities. This suggests that the workflow could be adapted to include a third "ambiguous" class for lower-probability classifications to be flagged for further manual review.

Table 3. Count of in/correct classifications by P(class = predicted)

Probability	TN	FN	TP	FP
>90%	677	3	24	1
(80, 90]	10	3	2	0
(70, 80]	4	0	0	4
(60, 70]	8	0	0	2
(50, 60]	8	0	0	4

Figure 5 shows results formed by varying "prediction level" between 0 and 100%. The ROC (receiver operating characteristic) curve shows minimal tradeoff between true and false positive rates. Given the unbalanced nature of the data set, the plot of precision vs. recall (P/R) is more meaningful than ROC for this example. The P/R curve shows that this model has skill (precision > 0.5) for most prediction levels, and that including the filter-tagged uninteresting (negative) cases from the test set when training the BN results in considerably better precision and recall than training the BN on the original training partition alone. Model performance depends strongly on prediction level (denoted by the labeled percentages in Figure 5), suggesting that tuning may be appropriate in practice.

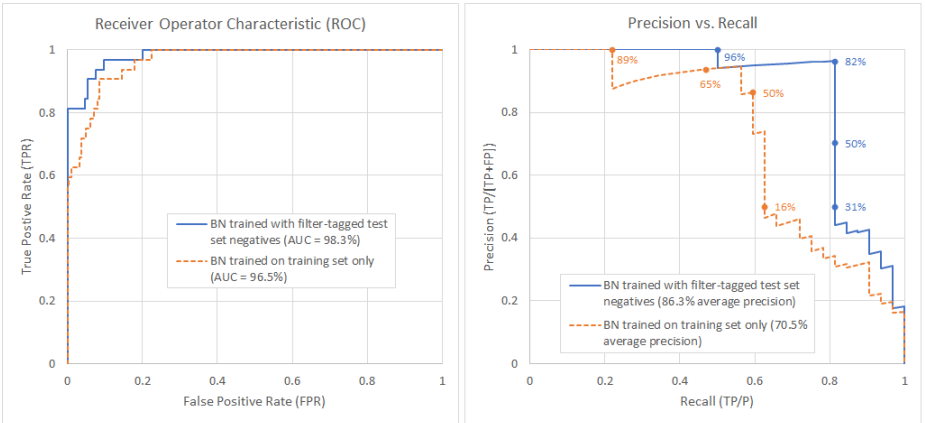

Fig. 5. ROC and Precision vs. Recall curves (varying prediction levels)

4 Conclusion and Future Work

We describe a parsimonious model for detecting rare events of interest from sparse, imbalanced data. The novelty filter allows fine control over the amount of data the SME must review. Using a Bayesian Network for classification allows detecting multivariate pattern differences between classes, enables partial learning from missing and untagged data, and natively provides probability estimates for classifications. Both the novelty filter and the Bayesian Network are explainable "glass box" methods whose results can readily be examined to understand why certain scores or classifications were provided – which we expect to be invaluable for human-in-the-loop interactive analytics. We show promising model performance on a synthetic data set designed to represent some of the challenges specific to detecting rare events of interest from small, sparse multivariate data.

With proof of concept demonstrated, performance comparison against other methods on a diverse range of datasets is prudent. While our method is intended to work only on nonnegative features with positive directionality, such features should be attainable from other data sets via appropriate transformation and feature extraction methods. Further study is warranted into use of classification probabilities in analysis (e.g., classifying samples as ambiguous). Investigation of other BN structures and filtering techniques may be appropriate for some data. Expansions of this method for data with temporal patterns is of interest and should be feasible via use of Dynamic BNs in combination with augmentation of the novelty filter analysis with features that encode state change detections and other temporal patterns.

Acknowledgements

This research was funded by the NNSA Office of Defense Nuclear Nonproliferation Research and Development, Office of Proliferation Detection (NA-221). Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. SAND No. SANDxxxx-yyyy J

References

1. Harrison, D.C., Seah W.K.G., Rayudu, R.: Rare event detection and propagation in wireless sensor networks. ACM Comput. Surv. 48, 4, Article 58, 22 pages. DOI: http://dx.doi.org/10.1145/2885508 (2016)
2. Zhao, J.H., Li, X., Dong, Z.Y.: Online Rare Events Detection. In: Zhou ZH., Li H., Yang Q. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2007. Lecture Notes in Computer Science, vol 4426. Springer, Berlin, Heidelberg (2007)

3. Benferhat, S., Tabia, K.: On the Detection of Novel Attacks Using Behavioral Approaches. ICSOFT 2008 - Proceedings of the Third International Conference on Software and Data Technologies, Volume PL/DPS/KE, Porto, Portugal, pp. 265-272. (2008)
4. Cheon, S-P., Kim, S., Lee, S-Y., Lee, C-B.: Bayesian networks based rare event prediction with sensor data. Knowledge-Based Systems, Volume 22, Issue 5, pp. 336-343, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2009.02.004 (2009)
5. Wong, W., Moore, A., Cooper, G., Wagner, M.: Bayesian network anomaly pattern detection for disease outbreaks. In T. Fawcett and N. Mishra, editors, Proceedings of the Twentieth International Conference on Machine Learning, pp. 808–815, Menlo Park, California, August 2003. AAAI Press (2003)
6. Uusitalo, L.: Advantages and challenges of Bayesian networks in environmental modelling. Ecological Modelling, Volume 203, Issues 3–4, pp. 312-318, ISSN 0304-3800, https://doi.org/10.1016/j.ecolmodel.2006.11.033. (2007)
7. Zheng, F., Webb, G.I.: Tree Augmented Naive Bayes. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA (2011)
8. Shi, H-B., Huang, H-K.: Learning tree-augmented naive Bayesian network by reduced space requirements. Proceedings of the International Conference on Machine Learning and Cybernetics, Beijing, China, pp. 1232-1236 vol.3. (2002)
9. Tjoa, E., Guan, C.: A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. arXiv:1907.07374v5 [cs.LG] (2020)
10. Vishnu, TV, Gugulothu, N., Malhotra, P., Vig, L., Agarwal, P., Shroff, G.: Bayesian Networks for Interpretable Health Monitoring of Complex Systems. In AI4IOT Workshop at International Joint Conference on Artificial Intelligence (IJCAI) (2017)
11. Wiegerinck W., Burgers W., Kappen B.: Bayesian Networks, Introduction and Practical Applications. In: Bianchini M., Maggini M., Jain L. (eds) Handbook on Neural Information Processing. Intelligent Systems Reference Library, vol 49. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-36657-4_12 (2013)