A New Frontier in Novelty Detection: Pattern Recognition of Stochastically Episodic Events

Colin Bellinger and B. John Oommen*

School of Computer Science, Carleton University, Ottawa, Canada
{cbelling,oommen}@scs.carleton.ca

Abstract. A particularly challenging class of PR problems in which the, generally required, representative set of data drawn from the second class is unavailable, has recently received much consideration under the guise of *One-Class* (OC) classification. In this paper, we extend the frontiers of OC classification by the introduction of a new field of problems open for analysis. In particular, we note that this new realm deviates from the standard set of OC problems based on the following characteristics: The data contains a *temporal* nature, the instances of the classes are "interwoven", and the labelling procedure is not merely impractical - it is almost, by definition, impossible, which results in a poorly defined training set. As a first attempt to tackle these problems, we present two specialized classification strategies denoted by Scenarios $S1$ and $S2$ respectively. In Scenarios $S1$, the data is such that standard binary and one-class classifiers can be applied. Alternatively, in Scenarios $S2$, the labelling challenge prevents the application of binary classifiers, and instead, dictates a novel application of OC classifiers. The validity of these scenarios has been demonstrated for the exemplary domain involving the Comprehensive Nuclear Test-Ban-Treaty (CTBT), for which our research endeavour has also developed a simulation model. As far as we know, our research in this field is of a pioneering sort, and the results presented here are novel.

Keywords: Pattern Recognition, Novelty Detection, Noisy Data, Stochastically Episodic Events.

1 Introduction

A common assumption within supervised learning is that the distributions of the target classes can be learned, either parametrically or non-parametrically. Moreover, it is assumed that a representative set of data from these classes is available for the training of supervised learning algorithms; indeed, the latter implies the former.

Beyond this, there exists a special form of Pattern Recognition (PR), which is regularly denoted *One-Class* (OC) classification [4,5,6,7,10,11]. This

* *Chancellor's Professor; Fellow: IEEE and Fellow: IAPR.* The Author also holds an Adjunct Professorship with the Dept. of ICT, University of Agder, Norway.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 435–444, 2011.
© Springer-Verlag Berlin Heidelberg 2011

"exceptional" category of binary classification is noteworthy in lieu of the fact that drawing a representative set of data to compose the second class (ω_2), is abnormally arduous, if not altogether impossible.

PR tasks of this nature have previously been constituted as involving outlier (or novelty) detection as the vast majority of the data takes a well-defined form that can be learned, and that samples from the ω_2 class will appear anomalously – outside the learned distribution. Although such problems represent a significant challenge, the results reported in the literature demonstrate that satisfactory results can often be obtained (see [4,5,6,7,10,11], for example).

In the subsequent section, Section 2, we introduce an advanced category of OC learning. Section 3, proceeds to draw a conceptual distinction between the target domain and those to which OC classifiers have traditionally been applied. The set of OC learners applied in this research are considered in Section 4. Section 5 describes an experiment based on the exemplary task of verifying the Comprehensive Nuclear Test-Ban-Treaty (CTBT). The results of the experiments, and a subsequent discussion, are contained in Section 6 and Section 7 respectively. Finally, Section 8 consists of our concluding remarks.

2 SE Event Recognition

To expand the horizon of the field, we observe that there exists a further, and yet more challenging subset of the OC classification domain of problems, which remains unexplored. We have denoted this class of problem as Stochastically Episodic (SE) event[1] recognition.

The problem of SE event recognition can be viewed in a manner that distinguishes it from the larger set of OC classification tasks. In particular, this category of problems has a set of characteristics that collectively distinguish it from its more general counterparts. The characteristics of this category can be best summarized as follows:

- The data presents itself as a time sequence;
- The minority class is challenging to identify, thus, adding noise to the OC training set.
- The state-of-nature is dominated by a single class;
- The minority class occurs both rarely and randomly within the data sequence.

Typically, in OC classification, the accessible class, and in particular, the data on which the OC classifier is trained, is considered to be well-defined. Thus, it is presumed that this data will enable the classifier to generalize an adequate function to discriminate between the two conceptual classes. This, for example,

[1] Events of this nature are denoted stochastic because their appearances in the time-series are the results of both deterministic and non-deterministic processes. The non-deterministic triggering event could, for example, be the occurrence of an earthquake, while the transmission of the resulting p- and s-waves, which are recorded in the time-serise, are deterministic.

was demonstrated in [5], where a representative set of the target computer user's typing patterns, which are both easily accessible and verifiable, were utilized in the training processes.

The classification of SE events is considerably more difficult because deriving a strong estimate of the target class's distribution is unfeasible due to the prospect of invalid instances (specifically members of the ω_2 class erroneously labelled ω_1) in the training set.

Under these circumstances, we envision two possible techniques for discriminating between the target class and the SE events of interest. The first scenario, denoted S1, involves application of standard clustering/PR algorithms to label both the classes appropriately. Alternatively, there are no instances of the ω_2 class available in S2, and the ω_1 class is poorly defined. Thus, novel applications of traditional OC classifier are applied.

3 Characteristics of the Domain of Problems

To accentuate the difference between the problems that have been studied and the type of problems investigated in this research, we refer the reader to Table 1. This table displays an assessment of six classification problems that have previously appeared in the literature on OC classification. In addition, we include the CTBT verification problem, which we present as a model SE event recognition problem. The first column indicates whether the problem has traditionally been viewed as possessing an important *temporal* aspect. The three entries with an asterisk require special consideration. In particular, we note that while traditionally these domains have not been studied with a temporal orientation, they do, indeed, contain a temporal aspect. The subsequent column signals whether the manual labelling of data drawn from the application domain is a significant challenge. This is, for example, considered to be a very difficult task within the field of computer intrusion detection, where attacks are well disguised in order to avoid detection.

Table 1. A comparison of well-known OC classification problems. The explanation about the entries is found in the text.

Dataset	Temporal	ID	Imbalance	Imbalance	Interwoven
		Challenge	*Type I*	*Type II*	
Mammogram	No	Low	Yes	Medium	No
Continuous typist recognition	No	Low	Yes	Medium	No
Password hardening	No	Low	Yes	Medium	No
Mechanical fault detection	No*	Low	Yes	Medium	No
Intrusion detection	No*	High	Yes	High	No
Oil spill	No*	High	Yes	Medium	No*
CTBT verification	Yes	High	Yes	High	Yes

The following two columns quantify the presence of class imbalance. In the first of these, we apply a standard assessment of class imbalance, one which relies on the determination of the *a priori* class probabilities. Our subsequent judgement departs slightly from the standard view, and considers class imbalance that arises from the difficulty of acquiring measurements (due to cost, privacy, *etc.*). The final column specifies if the minority class occurs rarely, and randomly (in *time* and magnitude), and if it occurs within a time sequence dominated by the majority class.

To summarize, in this section we have both demonstrated the novelty of this newly introduced sub-category of PR problems, and positioned the CTBT verification task within it. We additionally note that the fault detection, intrusion detection, and oil spill problems could be reformulated to meet the requirements of our proposed category. This, indeed, suggests a new angle from which these problems can be approached.

4 Classification

In all brevity, we mention that the binary classifiers used in this study were the Multi-layer Perceptron (MLP), the Support Vector Machine (SVM), the Nearest Neighbour (NN), the Naïve Bayes (NB) and the Decision Tree (J48), all of which are fairly well known, and so their descriptions are omitted here.

Alternatively, this work employed the following OC classification techniques: *a)* autoassociator (AA) [6], the Combined Probability and Density Estimator (PDEN) [5], one-class Nearest Neighbour (ocNN) algorithm [3], and the the scaled ocNN (socNN) [2].

Each of the applied classifiers has been implemented in the Weka machine learning software suite.

4.1 Classification Scenarios

Two possible SE event recognition problem exist. The S1 scenario assumes that through PR, or Clustering, means, all the instances of the minority class can be separated from the majority class for training purposes. Furthermore, the ω_2 class is large enough to explore binary classification.

In S2, however, the primary class will not be well-defined, as it is likely to contain erroneously labelled instances of the outlier class. This is a result of the impracticality of manually identifying and labelling them. In addition, the hidden minority class is extremely small.

For S2, we propose the application of OC classifiers in an unsupervised manner. In particular, they are trained on datasets in which the vast majority of instances have correctly been extracted from the background class. However, the impracticality of identifying the rare SE events implies the probable presence of some erroneous training instances.

We submit that by utilizing estimates of the state-of-nature, the problems associated with the noisy training set can be overcome through the appropriate parametrization of an internal *rejection rate* parameter.

The general performance of the classifiers are examined across all of the simulated detonation ranges. In addition, the performance as a function of distance is examined.

5 Experimental Setup

In this section, we present a series of experiments based on the verification of the CTBT. These experiments are designed to both illustrate the domain of SE events, and to exhibit a first attempt at SE events recognition.

5.1 Application Domain

The CTBT aims to prevent nuclear proliferation through the banning of all nuclear detonations in the environment. As a result, a number of verification strategies are currently under study, aimed at ensuring the integrity of the CTBT. The primary verification technique being explored relies on the quantity of radioxenon measured continuously at individual receptor sites, distributed throughout the globe. Radionuclide monitoring, in general, has been identified as the sole technique capable of unambiguously discriminating low yield nuclear detonations from the background emissions. More specifically, verification of the treaty based on the four radioxenon isotopes, ^{131}Xe, ^{133}Xe, ^{133m}Xe and ^{135}Xe, has been promoted due to the relatively low background levels, their ideal rates of decay and inert properties [8,9].

In general, the measured radioxenon levels are expected to have resulted from the industrial activities, such as nuclear power generation and medical isotope production. However, they are also the byproducts of low yield clandestine nuclear weapons tests, which are the subject of the CTBT.

5.2 Procuring Data: Aspects of Simulation

As a means of acquiring experimental datasets for this research, we utilized the simulation framework presented by Bellinger and Oommen in [1]. Their simulation framework models SE events, such as earthquakes, nuclear explosions, etc., as they propagate through the background noise, in this case representing radioxenon emitted from industry into the earth's atmosphere.

While it is generally beneficial to develop and study classifiers on "real" data, this is, indeed, impossible within the CTBT verification problem due to the absence of measured detonations, and the limited availability of background instances.

6 Results

In this section, we present the results that were obtained according to the four assessment criteria motivated in the previous sections, on the first classification scenario. We commence our exploration of PR performance by examining the AUC scores produced by each classifier over the 23 detonation ranges.

6.1 Results: Scenario 1

We first present the experimental results obtained for Scenario S1.

General Performance: With regard to the results, we include a general overview of the performance levels of each of the considered classifiers on the simulated CTBT domain. More specifically, we present an assessment of the five binary classifiers and the four OC classifiers, in terms of their AUC scores averaged over the 230 datasets that spanned the 23 detonation ranges.

In light of the fact that the SE events, which are to be identified, will, in practice, occur at random and unpredictable distances, these results yield a particularly insightful overview of the general performance levels.

The SVM classifier is, surprisingly, by far the worst performing classifier on this data, and in spite of its bias, it is, on average, worse than the OC classifiers, AA and socNN. This is demonstrates in Table 2, which contrasts the mean AUC scores of AA and socNN as 0.656 and 0.603, respectively, with the mean value for the SVM classifier of 0.528. Moreover, all four OC classifiers appear superior to the SVM when considered in terms of their maximum AUC scores.

The binary classifier, the MLP, stands out as the superior classifier, with J48, NN, and NB contending for the intermediate positions. The results posted in Table 2 confirm that the MLP is the strongest of the classifiers considered here. Furthermore, it indicates that the J48 and NB are very similar, and that the NN is the fourth ranking binary classifier according to the mean and maximum scores. However, the NN is second when ranked according to the minimum AUC scores.

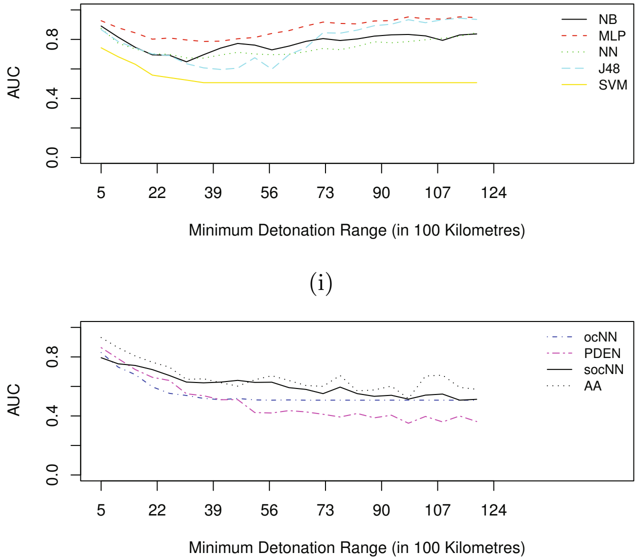
Table 2. This table displays the general classification results, in terms of AUC

	Mean	Max	Min	STDV
NB	0.772	0.939	0.504	0.074
MLP	0.869	0.976	0.674	0.067
NN	0.741	0.913	0.584	0.071
J48	0.774	0.98	0.500	0.148
SVM	0.528	0.813	0.500	0.065
ocNN	0.540	0.875	0.496	0.087
PDEN	0.487	0.943	0.182	0.156
socNN	0.603	0.842	0.405	0.094
AA	0.656	0.970	0.251	0.140

Notably, of the set of OC classifiers, PDEN produced the most variable range of the AUC scores. It is our suspicion that this variability resulted from the PDEN's generation of an artificial second class in its training process. However, further exploration of this matter is required. In general, the AA classifier is identified as the strongest OC classifier, both with respect to its mean and median values. While the socNN classifier achieved the second highest mean, it is more stable than the AA, with a lower standard deviation.

Performance as a Function of Distance: These results are particularly interesting, as they provide greater insight into performance trends, and suggest a performance scale for successively sparser receptor networks.

The performance plots depicted in Figure 1 reflect the ensemble mean of each classifier's performance at the 23 detonation ranges.



(i)

(ii)

Fig. 1. In this figure, plot (i) displays the performance of the five binary classifiers, in terms of their AUC scores, as a function of distance. Similarly, plot (ii) displays the performances of the four OC classifiers as a function of distance, according to their AUC scores.

Within Figure 1 (see Plot (i)), the MLP classifier is identifiably the superior classifier to the remaining four binary learners in terms of the AUC, across the range of detonation distances. In addition, it is not subject to the abrupt fluctuations that J48, and to a lesser extent NB, incur.

Plot (ii) in Figure 1 presents the performance of the OC learners. All of the OC classifiers follow a similar downward trend, which occurred between 0.8 and 0.9, towards, or beyond in the case of the PDEN, an AUC of 0.5. The AA and the socNN degrade in a slower, and in a more linear fashion than PDEN and ocNN.

6.2 Results: Scenario 2

In this section, we explore the very intriguing classification scenario S2. More specifically, we present an assessment of the four OC classifiers, in terms of their AUC scores on the 230 datasets that covered the 23 detonation ranges.

General Performance: We first present a general overview of the performance of the set of OC classifiers on the simulated CTBT domain.

In light of the fact that the SE event will, in practice, occur at random and unpredictable distances, these results are particularly insightful.

Table 3 contains a compilation of the mean, maximum, minimum and standard deviation of the each classifier's overall results.

Table 3. This table displays the general classification results, in terms of AUC

	Mean	Max	Min	STDV
ocNN	0.505	1	0.496	0.042
PDEN	0.507	1	0.075	0.185
socNN	0.587	1	0.292	0.171
AA	0.621	1	0.024	0.225

Our assessments of Table 3 reveal that, similar to our findings on the S1 scenario, the AA classifier is superior, in terms of its mean, and median scores, to the other OC classifiers. Indeed, on this, which is a more challenging task, its mean and median values are only slightly lower than in the previous task. However, within this second scenario, it has the lowest minimum AUC scores. It is also extremely unstable, with results ranging from perfect to near zero.

The classifier, socNN, ranks second after the AA according to its mean, and was considerably more stable, while the ocNN and PDEN classifiers produced values that were near or below 0.5.

Performance as a Function of Distance: As in the case of Scenario S1, we have also studied the performance of the classifiers as a function of distance, where the latter is assessed according to the AUC.

The AA and socNN are, once again, roughly identifiable as the best of the four classifiers in Figure 2. However, all of the classifiers, with the exception of ocNN, which rapidly converges to 0.5, suffer from significant and essentially random fluctuations. These fluctuations in performance suggest that the classifiers' results were as dependent on the nature of the SE events in the 230 datasets, as on the distance at which the events originally occurred.

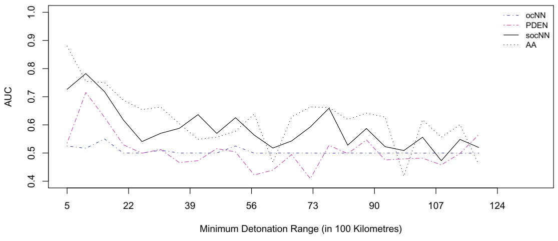


Fig. 2. This figure displays the performance of the four OC classifiers as a function of distance, according to their AUC scores

7 Discussion

The relatively low mean AUC scores produced by the OC classifiers, and variability in their results on the CTBT feature-space, clearly illustrate the many challenges in the application of OC learners. However, we suspect that the results are not so imbalanced, as Hempstalk *et al.*, in [5], noted that such comparisons are generally biased towards binary learners.

The initial performances of the OC classifiers suggests that they are very capable of associating anomalously high levels of radioxenon with the SE event class. However, the binary learners are not only well adapted to classifying anomalously highly levels as members of the SE event class, they are also capable of classifying anomalously low levels, which commonly result from detonations that occurred well beyond the radial distance to the background source and advected from a different direction.

The instability in performance that is depicted with respect to distance that appears in S2, results both from the erroneous instances in the training sets, and the variability in the classification challenges presented by the few members of the SE event class in the test sets. Indeed, the generation of random SE events over a domain as vast as the simulated CTBT domain, will inevitably produce both very easy, and nearly impossible classification tasks. Thus, when randomly including only a minute number of these events in the test sets, it is probable that performance on the SE event class will fluctuate significantly. This is, of course, why a large number of receptors are required in the global receptor network.

However, while the ensemble mean performance fluctuates considerably over the successive experiments, when considered in terms of the overall means, the performance of the OC classifiers on the S2 task is only slightly lower than on the S1 task. This is, indeed, a promising result.

8 Conclusion

In this research, we have extended the frontiers of novelty detection through the introduction of a new field of problems open for analysis. In particular, we noted that this new realm deviates from the standard set of OC problems based on the presence of three characteristics, which ultimately amplify the classification challenge. They involve the *temporal* nature of the appearance of the data, the fact that the data from the classes are "interwoven", and that a labelling procedure is almost, by definition, impossible.

As a first attempt to tackle these problems, we presented two specialized classification strategies as demonstrated within the exemplary scenario intended for the verification of the Comprehensive Nuclear Test-Ban-Treaty (CTBT). More specifically, we applied the simulation framework presented by Bellinger and Oommen, in [1], to generate CTBT inspired datasets, and demonstrated these classification strategies within the most challenging classification domain. More specifically, we have shown that OC classifiers can successfully be applied to classify Stochastically Episodic (SE) events, which are unknown, although present, at the time of training.

The problem of including the temporal aspects of SE events in a PR methodology (for example, by invoking a time series analysis) remain open.

References

1. Bellinger, C., Oommen, B.J.: On simulating episodic events against a background of noise-like non-episodic events. In: Proceedings 42nd Summer Computer Simulation Conference, SCSC 2010, Ottawa, Canada, July 11-14 (2010)
2. Bellinger, C., Oommen, B.J.: Unabridged version of this paper (2010)
3. Datta, P.: Characteristic concept representations. PhD thesis, Irvine, CA, USA (1997)
4. Ghosh, A.K., Schwartzbard, A., Schatz, M.: Learning program behavior profiles for intrusion detection. In: Proceedings of the Workshop on Intrusion Detection and Network Monitoring, vol. 1, pp. 51–62 (1999)
5. Hempstalk, K., Frank, E., Witten, I.H.: One-class classification by combining density and class probability estimation. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part I. LNCS (LNAI), vol. 5211, pp. 505–519. Springer, Heidelberg (2008)
6. Japkowicz, N.: Concept-Learning in the Absence of Counter-Examples: An Autoassociation-Based Approach to Classication. PhD thesis, Rutgers University (1999)
7. Kubat, M., Holte, R.C., Matwin, S.: Machine learning for the detection of oil spills in satellite radarimages. Machine Learning 30(2), 195–215 (1998)
8. Saey, P.R.J., Bowyer, T.W., Ringbom, A.: Isotopic noble gas signatures released from medical isotope production facilities – Simulation and measurements. Applied Radiation and Isotpes (2010)
9. Stocki, T.J., Japkowicz, N., Li, G., Ungar, R.K., Hoffman, I., Yi, J.: Summary of the Data Mining Contest for the IEEE International Conference on Data Mining, Pisa, Italy (2008)
10. Tarassenko, L., Hayton, P., Cerneaz, N., Brady, M.: Novelty detection for the identification of masses in mammograms. In: IEE Conference Publications 1995(CP409), pp. 442–447 (1995)
11. Tax, D.M.J.: One-class classification; Concept-learning in the absence of counter-examples. PhD thesis, Technische Universiteit Delft, Netherlands (2001)