On the Pattern Recognition and Classification of Stochastically Episodic Events*

Colin Bellinger[1] and B. John Oommen[2]

[1] The School of Information Technology and Engineering,
University of Ottawa, Ottawa, Canada
cbell052@uottawa.ca
[2] *Chancellor's Professor*, *Fellow: IEEE* and *Fellow: IAPR*
School of Computer Science,
Carleton University, Ottawa, Canada
oommen@scs.carleton.ca

Abstract. Researchers in the field of Pattern Recognition (PR) have traditionally presumed the availability of a representative set of data drawn from the classes of interest, say ω_1 and ω_2 in a 2-class problem. These samples are typically utilized in the development of the system's discriminant function. It is, however, widely recognized that there exists a particularly challenging class of PR problems for which a representative set is not available for the second class, which has motivated a great deal of research into the so-called domain of One Class (OC) classification. In this paper, we extend the frontiers of novelty detection 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 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 not merely impractical - it is almost, by definition, impossible. 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 the novel application of one-class 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, Rare Events, Stochastic Events, Erroneous Data.

* The first author is also an *Adjunct Professor* with the University of Agder in Grimstad, Norway. Both the authors are grateful for the partial support provided by NSERC, the Natural Sciences and Engineering Research Council of Canada. A preliminary version of this paper was presented at ACIIDS'11, the 2011 Asian Conference on Intelligent Information and Database Systems, in Daegu, Korea, in April 2011.

N.T. Nguyen (Ed.): Transactions on CCI VI, LNCS 7190, pp. 1–35, 2012.

1 Introduction

1.1 Problem Formulation

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 commonly-reported method of classification, there exists a special form of Pattern Recognition (PR), which is regularly denoted One Class (OC) classification [10,12,14,16,30,31]. This "exceptional" category of binary classification is noteworthy in lieu of the significant challenge that it presents. Escalating the difficulty, is the fact that drawing a representative set of data to compose the second class (ω_2), which is fundamental to the derivation of a binary discriminant function, is abnormally arduous, if not altogether impossible. The difficulty of acquiring a sufficiently symbolic set may arise because of:

1. The natural *imbalance* in the classification task;
2. The difficulty (due to cost, privacy, etc.) of acquiring samples from the ω_2 class;
3. The task of obtaining representative samples of the ω_2 class is overwhelming, as a result of the vastness of the distribution.

PR tasks of this nature have previously been constituted as involving outlier (or novelty) detection in lieu of the fact that the vast majority of the data takes, what is assumed to be, 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 can be significantly more difficult than those that involve two well-defined classes of data, the results reported in the literature demonstrate that satisfactory results can often be obtained (see [10,12,14,16,30,31], for example).

1.2 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 problems as Stochastically Episodic (SE) event 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 problem 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 unwarranted noise to the one-class 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 PR solutions to so-called OC problems, 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, was demonstrated in [30], where the training set consisted exclusively of images of non-cancerous tissue. Similarly, in [12], 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[1] 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. In this work, we present solutions to this problem based on tradition one-class classifiers.

SE event recognition is additionally challenging because the validity of instances drawn from the target class are suspect, and the occurrences of the minority class are temporally (i.e. with respect to the time-axis) interwoven with the data from the majority class.

1.3 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 one-class classification problems, which, while only a small subset, cumulatively illustrate the traditional scope of the problem set. In addition, we include the problem of CTBT verification, which forms our exemplary SE event recogonition 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 subvert the system.

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

[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.

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

Dataset	Temporal	ID Challenge	Imbalance *Type I*	Imbalance *Type II*	Interwoven
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

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 (briefly) 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.

1.4 Overview of Our Solution

As previously indiated, SE event recognition composes a particular challenging problem due to the combined affect of the four characteristics that are inherent in such problems. Under these circumstances, we envision two possible techniques for discriminating between the target class and the stochastically episodic events of interest. If the incoming training data contains a sufficient quantity of accurately identifiable stochastic events, a standard clustering/PR algorithm could be applied to label both the classes appropriately. Subsequent to the labelling procedure, a standard binary classifier could be trained and utilized to achieve the classification of novel instances. In this body of work, we refer to this scenario as S1, and the subsequent scenario as S2.

Alternatively, and more applicable in scenarios in which the SE events are extremely rare, all of the training data can be assigned to the target class, and an OC classifier can be applied. The details of, and justification for, this approach are described in the subsequent sections. Our primary objective in this research is to illustrate how standard supervised learning algorithms can be applied to discriminate rare stochastic episodes, which apart being unanticipated, are random in magnitude and position within the sequence of background data.

1.5 Contributions of This Paper

The novel contributions of this paper, with respect to PR, are as follows:

- We introduce an important new category of PR, namely SE event recognition. In particular, we note that this new realm deviates from the standard set of one-class problems based on the presence of four characteristics: *(a)* the data presents itself as a time sequence; *(b)* the minority class is challenging to identify, thus, adding unwarranted noise to the OC training set; *(c)* the state-of-nature is dominated by a single class; and, *(d)* the minority class occurs both rarely and randomly within the data sequence.
- In addition, we present a first attempt at classifying SE events within the examplary verification problem suggested by the Comprehensive Test-Ban-Treaty (CTBT). Our initial approach is extremely accessible, as it is based on "off the shelf" PR solutions.
- More specifically, where the ω_2 is sufficiently large, we demonstrate how clustering/PR algorithms can be applied to label training data for the development of a sound binary classifier.
- Finally, in scenarios where training instances cannot be acquired from the second class (the so-called OC problem), and where the accessible class in known to contain noise due to labeling issues, we illustrate how, through novel means, standard OC classifiers can be applied as unsupervised learners.

We conclude this section by mentioning that our results probably represent the state-of-the-art!

1.6 Paper Organization

The rest of the paper is organized as follows. In Section 2 we presented a brief survey of the available solutions for dealing with PR. Subsequently, in Section 3, we present the application domain, and our solution to modelling SE event systems for the purpose of PR system development. Then, in Section 4 we present a brief overview of issues of PR in relation to SE event recognition. Thereafter, experimental results obtained by rigorously testing our solution on the exemplary scenarios suggested by the CTBT are presented in Section 5 and 6, and discussed in Section 7. Section 8 concludes the paper.

2 Pattern Recognition: State of the Art

This section[2] serves to present the state-of-the-art in PR. In that regard, Duda, *et al.*, in [9] describe pattern recognition as follows:

> "The act of taking in raw data and taking an action based on the 'category' of the pattern."

[2] This brief section has been included in the interest of completeness. Although these issues are considered commonplace for the general PR problem, they are still fairly non-standard for OC problems - which advocates the necessity of the section.

It is, indeed, natural that we should desire to 'teach' machines to recognize sets of patterns that are easily recognizable to humans, such as handwritten characters, speech and faces, as computers present the possibility of increased efficiency and do not become tired of mundane tasks. Furthermore, the benefits of training machines to classify complex patterns, typically left to doctors and scientists with considerable specialization in the domain, are equally apparent. Thus, researchers have continued to push the state-of-the-art in PR systems since the advent of the modern computer.

Supervised Learning. Prior to application, the PR system must be trained to discriminate between the objects of interest in its particular application domain. For multi-class problems, such as discrimination between handwritten characters, the PR system is said to learn a mapping that discriminates between the individual inputs by directing them to their corresponding categories. Alternatively, in the special scenario, which is of primary interest in this work, termed OC learning, instances of a single target category are available for the training of the PR system. As a result, the system takes a recognition-based approach, and attempts to learn a function that maps novel instances of the target category to the target class, and all others to the outlier class.

Broadly speaking, standard PR systems for supervised learning are trained on datasets drawn from their prospective application domains, in which each feature vector has been accented with its corresponding class label. The objective of the training process is the derivation of a set of models that articulate the individual characteristics of the classes. Thus, while the performance on the training set is of little interest, rather, the focus shifts to the selection of a model that will perform well on novel instances in the future. The derivation of these models is algorithm-specific, however, there exists commonalities between all learners. Generally speaking, regardless of the learning strategy, the accuracy of the derived model on novel instances will increase with the size of the training set. In addition, all learners strive to optimize the balance between specialization and generalization [18].

Under ideal circumstances, the training procedure for a binary learner is able to rely on an ample supply of data that has been uniformly drawn from both classes. As a result, increasingly accurate models of the classes in question can be constructed, and therefore, an effective classifier of novel instances is produced. Conversely, this assumption cannot be made in SE event recognition problems. Thus, particular expertise is required in order to derive an acceptable model.

OC learning problems characteristically involve scenarios in which the available class is easily acquired and exists in abundance, while the second class is exceptionally difficult to acquire, or naturally rare [31]. During extreme class imbalance, the majority class can be expected to compose as much as ninety-five percent of the data. In such scenarios, it is typical that the class we are most interested in identifying is the minority class, as is the case in automated mammogram scans and many other medical disciplines [30]. Japkowicz, in [14], and Kubat *et al.*, in [16], demonstrate scenarios in which acquiring instances is both difficult and expensive. In particular, the challenge of Kubat *et al.* requires

the hand-labelling of satellite imagery, while the former involves fault detection in helicopter gearboxes, which are expensive to run. Moreover, the derivation of the outlier class would require the destruction of the gearboxes in an infinite number of ways. Alternatively, under certain conditions, the second class might be so large as to render the accumulation of a sufficient supply a seemingly insurmountable challenge. This scenario is well illustrated by the continuous typist recognition problem described by Hempstalk *et al.*, in [12]. The objective of the depicted classification challenge is to distinguish the sole legitimate terminal user from all other users. A proper training set, therefore, would be drawn uniformly from the set of all people, which is clearly infeasible.

A variety of approaches have been applied to OC classification. The more traditional of these involve extensions to existing binary classifiers or density estimations. The density estimation approach fits a statistical distribution, such as Gaussian, to the target data, and classifies novel instances based on the learned probability of their occurrences. Such a technique has been applied by the authors of [4,21,30]. Techniques that extend existing classifiers typically modify the inner structure of the classifier to fit boundaries around the target class, and classify those novel instances falling outside the boundary as outliers, as is demonstrated by [14,27]. These two approaches, in addition to some alternative approaches to one-class learning, such as the work described in [12], which is a combination of these two techniques, are discussed in the sections to follow.

Density Estimation. Density estimation is, perhaps, the most elementary of all approaches to OC classification. The fundamental idea behind this OC classification technique is the estimation of a Probability Density Function (PDF), $\hat{P}(\mathbf{x})$, based on a training set, $\mathbf{D}^n = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, drawn independently and identically from the underlying distribution, $P(\mathbf{x})$, of the target class. Subsequent to the estimation of the PDF, novel instances are classified according to a predefined target threshold or by resorting to a suitable statistical test.

Under ideal circumstances, and in particular, where sufficient training data is accompanied by a substantial understanding of the background distribution, or a flexible density estimation technique, density estimation-based classifiers are known to produce strong results [31]. However, a significant quantity of training data is required to overcome the curse of dimensionality, as is described by Duda *et al*, in [9].

Bishop, in [5], discusses three approaches to PDF estimation; the first of these techniques requires the modeller to provide an initial specification of the functional form of the underlying distribution, such as Gaussian or Poisson. An iterative process based on the predefined distribution, is applied to fit the density function to the training data through the optimization of the corresponding PDF parameters. The application of the parametric method is significantly limited by the fact that, in many cases, the specified PDF may be incapable of describing the training data.

Non-parametric estimation techniques represent a more flexible approach, as they do not assume a particular functional form, and instead allow the training data to completely specify the PDF. As a result, the PDF is not limited to a small

set of standard distributions, and does not have to be provided at initialization. However, the fact that the number of parameters to be optimized expands quickly as the dataset increases in size, can prove to be prohibitive.

Yeung and Chow, in [33], applied the non-parametric method for probability density estimation, introduced by Parzen, in [19], to the development of an intrusion detection system. More specifically, their approach utilized the Parzen-window estimation of $\hat{P}(\mathbf{x})$, with a Gaussian kernel, on a dataset composed of normal network activities. The generalized Parzen-window estimation of $\hat{P}(\mathbf{x})$, based on an n element dataset $\mathbf{D}$, takes the following form:

$$\hat{P}(\mathbf{x}) \;=\; \frac{1}{n} \sum_{i=1}^{i} \delta_n(\mathbf{x} - \mathbf{x}_i), \tag{1}$$

where $\delta_n(\cdot)$ is the kernel function (in this case, Gaussian in form), the exact form of which depends upon the number of instances in the training set. Subsequent to the training process, novel instances are classified based on their log-likelihood. In addition, the Parzen-window approach was previously applied by Tarassenko *et al.*, in [30], to the classification of anomalous mammograms.

A final approach, sometimes referred to as semi-parametric estimation, attempts to strike a balance between the previous two methods. This approach enables a general class of functional forms, in which the number of adaptive parameters is increased systematically to build a progressively more flexible model.

The mixture of Gaussians approach is a particular category of semi-parametric estimation schemes, which has received considerable application, as it is analytically attractive. This approach to semi-parametric estimation was applied in [22,23] to the detection of novel instances in a series of medical datasets, and as a procedure for noise removal in an image processing task. In its essence, the mixture of Gaussians method is composed of a linear combination of j Gaussian distributions, each of which is uniquely parametrized according to its particular mean, μ_j, and covariance, $\mathbf{\Sigma}_j$, such that

$$\hat{p}_{MoG}(\mathbf{x}) \;=\; \frac{1}{N_{MoG}} \sum_{j} \alpha_j \; pN(\mathbf{x}; \mu_j, \mathbf{\Sigma}_j), \tag{2}$$

where the α_js are the mixing coefficients.

2.1 One-Class Extensions to Binary Classifiers

Autoassociator. An autoassociator in an example of a feedforward Artificial Neural Network (ANN). However, unlike its more prevalent binary counterpart, the Multi-Layer Perceptron (MLP), which aims to produce a classification decision at the output layer, the autoassociator is trained to *reconstruct* the input vector at the output layer [24]. The general architectures for both forms of ANNs are illustrated in Figure 1.

The theoretical basis for the autoassociator relies on the fact that it is trained to compress and decompress instances of the concept class exclusively. Thus,

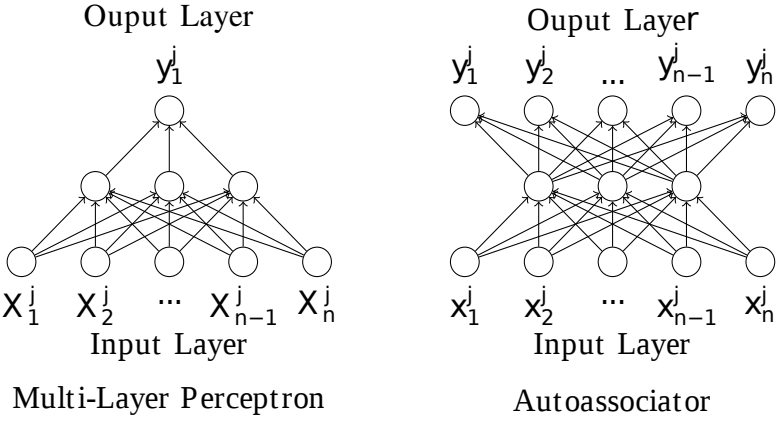

Fig. 1. This figure demonstrates two possible feedforward artificial neural network architectures. Subfigure (a) illustrates the general form of the Multi-Layer Perceptron (MPL). In Subfigure (b), the essential structure of an autoassociator is displayed.

during application, novel instances of the concept class should compress and decompress successfully. More specifically, the reconstruction error resulting from a novel member of the concept class, during application, is expected to be small. Alternatively, non-members of the target class are characterized by large reconstruction errors. Therefore, the classification procedure entails a comparison of the reconstruction error and a user-defined threshold. All instances reproduced with an error less than the threshold are considered to be members of the concept class, while the remainder are labelled as outliers, or non-members.

The OC classifiers of the above form have been applied in a number of domains with considerable success. Hanson and Kegl, in [11], introduced an autoassociator system, namely PARSNIP, developed to reconstruct syntactically correct sentences using the backpropigation procedure described by Rumelhart *et al.*, in [24]. The PARSNIP system, trained on the Brown University Corpus of Present-Day American English, in which the words of each sentence are tagged with their active syntactic category, learned to accurately identify sentences that were syntactically correct and reject those that were incorrect.

Subsequently, Petsche *et al.*, in [20], developed a system similar to a fuel gauge based on the principle of the autoassociator. The system described in that work, learned to predict the impending failure of a motor. Its intended application domain is characterized by a high cost associated with failures, such as the fire pump on navy vessels.

More recently, Japkowicz, [14], examined the performance of the autoassociator in comparison with a variety of binary learners on three domains. In particular, the case studies utilized a CH46 Helicopter gearbox dataset, with the objective of predicting the failure of the gearbox based on vibration time signals, and the sonar and DNA promoter datasets from the U.C. Irving Repository of Machine Learning. The recognition task in the former was to distinguish mines

from rocks in sonar data, while the objective in the DNA promoter dataset was to classify promoters in a DNA sequence. The autoassociator was found to be robust relative to the other classifiers in all three case studies, and more accurate on both the helicopter gearbox and DNA promoter tasks.

One-Class Support Vector Machines. Schölkopf *et al.*, in [28], proposed a one-class extension to the existing support vector techniques, for the estimation of support in high-dimensional spaces. In general terms, their approach maps the training data into a dot product feature space, and inserts a hyperplane in a manner that separates the origin from the data with maximal margin.

In their work on one-class SVMs, Schölkopf *et al.* explored their implementation on both artificial and real-world data. For the latter category, they used the US Postal Service's handwritten digits.

The handwritten digits dataset was converted to facilitate two distinct sets of experiments. In the first experiment, tests were conducted that specified a random set of instances drawn from a single class that composed the target data, and left all instances from the remaining nine classes to form a large set of outliers. On this experiment, the one-class SVM was found to correctly identify the target class 91% of the time, and had a false positive rate of 7%.

In the second experiment, ten binary features were added to the handwritten digits dataset; one new feature for each of the possible digits. These features were included to identify the class to the classifier during training, with the notion that the classifier would learn to recognize what each digit should look like. For this experiment, the OC classifier was trained on instances drawn from each class, with the additional features. The authors found that the OC SVM learned to accurately identify anomalous patterns, and the erroneously labelled instances.

Similar implementations of the OC SVM have subsequently been applied to a large number of problems. Manevitz and Yousef, in [17] for example, applied the OC SVM to discover text documents of similar topics to those in the training set, and compared the results to a set of alternate OC classifiers. They concluded that, while the OC SVM is very sensitive to parametrization, with the right parameter set, it outperformed the other classifiers considered in the study, with the exception of the OC ANN.

Further examples of previous applications of OC SVMs are to classify yeast gene regulation predictors in [15], and for image retrieval in [6].

2.2 Nearest Neighbour

The standard Nearest Neighbour (NN) algorithm is a binary classifier that takes a non-parametric approach to PR. More specifically, in its simplest form, the training process involves "remembering" all of the training instances and their corresponding labels. During application, a novel instance $\mathbf{x}$ is classified according to a majority vote rule, in which the k (k is an odd number specified by the user) NNs of $\mathbf{x}$, in the training set, are polled for their respective classes.

The novel instance is subsequently assigned to the class that is occupied by the majority of its neighbours [9].

The NN algorithm has seen considerable application. Horton and Nakai, for example, compared the NN to Naïve Bayes and to decision trees, in [13], on the problem of predicting the cellular localization sites of proteins in yeast and ecoli. In their study, Horton and Nakai reported favourable results for the NN classifier.

More recently, modifications have been made to the NN classifier to facilitate OC classification. Datta, in [7], adapted the standard NN algorithm to preform OC classification through the utilization of a threshold learned during the training phases. More specifically, the algorithm searches the training set for the pair of NNs that are separated by the greatest distance, which is denoted by τ. When classifying a novel instance, the distance between it and its NN is compared to the learned parameter. If the distance is less than or equal to τ, the novel instance is assigned to the positive class. Otherwise it is assigned to the negative class.

The author applied this implementation of the OC NN (ocNN) algorithm to a number of UCI datasets and found it to be comparable with other OC classifiers. It was additionally found to be comparable with the binary C4.5 decision tree classifier on some classes of the Breast Cancer Wisconsin, Pima Indian Diabetes and Wine domains.

Tax, in [31], provided a comprehensive survey of the performance of one-class classifiers on a number of artificial domains, in which an alternate adaptation of the standard NN algorithm for one-class learning was included. Notably, he identified the OC NN algorithm to be a poor performer in a general analysis of robustness against outliers. This is, indeed, a problem that we noted when applying the ocNN to the task of SE event recognition, due to the fact that the mislabelled instances of the ω_2 class in the training set often appear as outliers, which should not be generalized into the model of the background (ω_1) class.

Scaled Nearest Neighbour. In recognition of the limitations of ocNN within the domain of SE event recognition, we propose a more suitable NN classifier [2] – which is one of the contributions of this paper. This modification was motivated by the second classification scenario, namely the one referred to earlier as S2. In particular, this scenario is characterized by a series of rare SE events where:

- The data exists as a time-series;
- The state-of-nature is dominated by a single class (the ω_1 class composes more than, for example 90% of the instances);
- The minority class is nearly impossible to manually identify. Thus, it naïvely takes the ω_1 label *even in the training set.*

In this scenario, we have stated that due to the rarity of the outlier class, and the extreme challenge of manually labelling those instances in the training set, it can be naïvely issued to a OC classifier. Moreover, this can be done with considerable confidence, provided that an estimate of the *a priori* probability of the outlier class can be acquired. This hypothesis relies on the availability

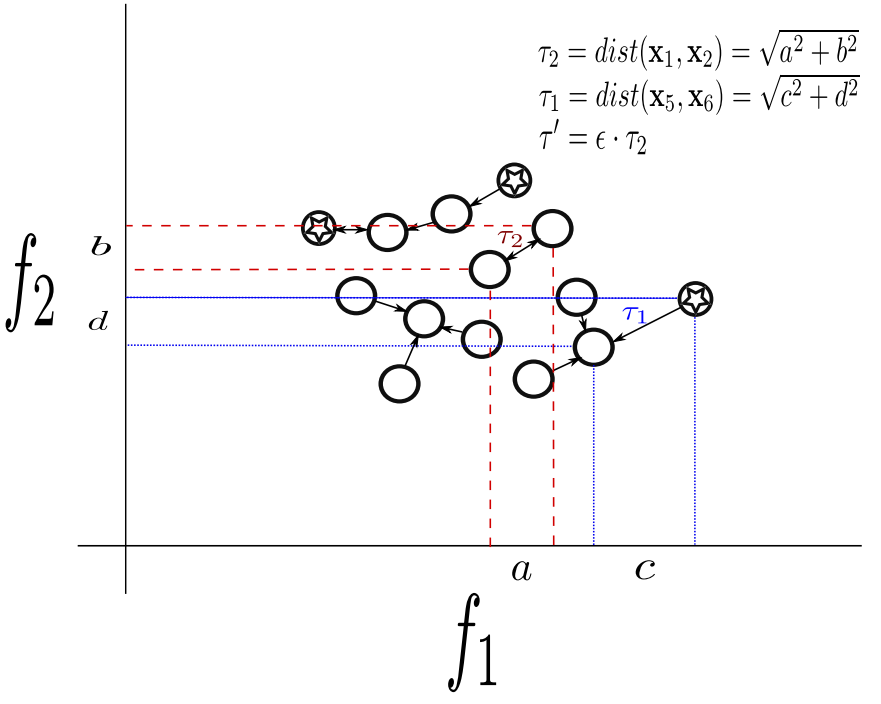

Fig. 2. This figure demonstrates the calculation of the τ parameter in the ocNN classifiers, and the effect of erroneous instances in the training set on the learned target rejection rate threshold

of a so-called rejection rate, which ensures a portion of the training set will be misclassified after the derivation of the discriminant function.

Observe that the standard ocNN algorithm is intuitively unable to learn a threshold capable of discriminating most of the erroneously labelled outliers, and is inherently ineffective in the presence of noise. The problem, which is embedded in the ocNN algorithm, is depicted in Figure 2. By definition, the naïvely labelled instances of the second class are outliers. Thus, they are expected to reside on the periphery of the "real" background distribution. Therefore, with a high probability, the learned parameter, τ, which is intended to record the variability in the background class [7], can be expected to represent the distance between a background instance and an erroneously labelled member of the outlier class. A hypothetically learned distance of this sort is illustrated in Figure 2 as τ_1. Ideally, however, the algorithm should learn the distance that is denoted as τ_2, because it is the maximum target rejection rate threshold found in the set of pure background instances (represented as empty circles).

Because this scenario creates an unsupervised learning environment, in which we cannot explicitly identify the members of the outlier class during training, we rely on a rejection rate parameter to be "engrained" in the OC classification

algorithm in order to facilitate the exclusion of these instances. However, while Datta coined τ to be the target rejection rate threshold, by definition, it does not exclude any instances in the training set. Indeed, this was not the intention. Thus, in this exceptional domain, it incorporates the erroneous information provided by the mislabelled members of the outlier class into the learned threshold, as is depicted by τ_1 in Figure 2.

As a means of accounting for the overestimate, we have added a scaling parameter, ϵ, where $0 < \epsilon \leq 1$, such that

$$\tau' = \epsilon \cdot \tau, \tag{3}$$

with the understanding that the optimal value of ϵ will enable the rejection of the majority of the outlier instances, by reducing the magnitude of the learned threshold.

2.3 Combined Density and Class Probability Estimation

Hempstalk *et al.*, in [12], introduced a technique for converting OC classification problems into binary tasks, based on a two-fold strategy. The initial phase of the strategy involves an examination of the training data for the concept class in order to determine its distribution. This knowledge is subsequently utilized in the generation of a non-concept, or outlier, class. In the second phase, a standard binary classifier is trained based on the concept class and the generated class. Most standard classification techniques are applicable here. The single limiting factor in the selection of a binary classifier is the requirement that the classifier of choice can produce a class probability estimate at prediction time. Using Bayes' rule, the authors demonstrate how the class density function can be combined with the class probability estimate to yield a description of the concept class.

The performance of the combined density and class probability estimation technique was examined on a multitude of datasets, the bulk of which result from the U.C. Irving Repository of Machine Learning. In addition, the performance was gauged on the very interesting task of recognizing a "continuous typist". This latter application required the validation of individual computer terminal users based on their learned typing patterns. With considerations founded upon these experiments, the authors concluded that the combination of the density function with a classification model can produce an improvement in accuracy beyond that which resulted from the density function or the classification model alone.

3 Modelling the Problem

To this point, we have described a novel sub-category of PR, which is characterized by the detection of a minute number of SE events interwoven in a time-series. Indeed, a number of interesting PR problems fit this form, including advanced earthquake, tsunami and machine failure warning systems, to name but a few. 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.

3.1 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 their inert properties [25,29].

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

3.2 Procuring Data: Aspects of Simulation

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. It has, however, been demonstrated that artificial data can be utilized for PR system development, and to generate controlled experiments (generalized case-studies), in the absence of "real" measurements [1,8]. In this vein, as a means of acquiring experimental datasets for this research, we utilized the simulation framework presented by Bellinger and Oommen in [3]. 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 the industry into the earth's atmosphere.

Simulation Scenario. In order to explore the PR of low yield clandestine nuclear tests, we devised a simulation scenario to capture the effects of a diverse set of detonation possibilities, within a realistic background scenario. In particular, and accordance with the majority of the CTBT's International Monitoring Station (IMS), the IMS in the simulated environment was impacted by a single industrial emitter. In this simulation scenario, the industrial emitter was positioned 3,000 km away from the IMS. Thus, when the atmospheric conditions transported the emitted radioxenon directly from the source to the receptor, and when the conditions were not conducive to the dispersion of the radioxenon, the background concentration could reached significant levels. However, due to the realistic atmospheric conditions that were built into the model, such as the fluctuations in wind speed and direction, along with atmospheric stability, the background levels were generally low. This fact is displayed by the histogram in Figure 3. The figure specifically demonstrates that the majority of the ^{131}Xe

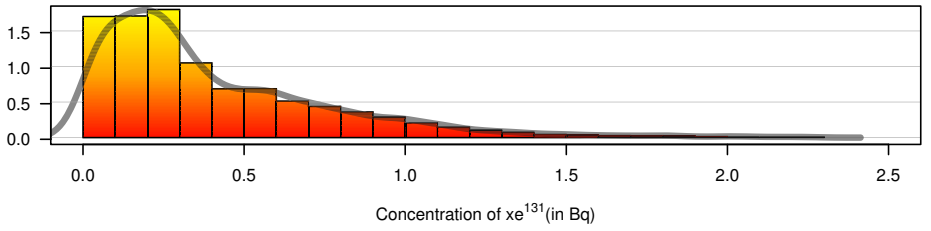

Fig. 3. This figure displays a histogram of the measured concentrations of ^{131}Xe at the IMS, resulting from the background source during the simulation

concentrations measured at the IMS site during the simulation were less than 0.5 $Bq\ m^{-3}$.

It is, however, highly probable that a clandestine detonation will occur at distances beyond the industrial source, thus, causing no, or only a minute, change in the radioxenon concentrations measured at the IMS, depending on the angular direction to the detonation site, and the prevailing meteorological conditions. Therefore, the classification of this type of SE event is extremely challenging.

With the above fact in mind, we considered the performance's of the PR systems as a function of distances. This is to specifically assess the probability of detecting detonations at various distances. In particular, 23 subcategories of datasets were generated. In each case, the modelled environment contained the same industrial source and IMS at the receptor site. As a result, for each simulation the background readings can be assumed to follow the distribution displayed in Figure 3. The 23 subsets formed a series of incremental detonation ranges, which commenced with all detonations occurring between 500 km and 1000 km, as illustrated in Figure 4.

The detonation range was iteratively increased by 500 km for each successive set. This incremental approach enabled the examination of performance as a function of distance, in addition to the more general considerations of performance.

As a binary classification problem, the generated sets were composed of two classes, in this case a background class and a detonation class. In addition to the class label, each instance was composed of the concentrations of the four isotopes measured by the IMS at the receptor site over the period of an hour. The simulation system contains two phases, the first phase simulates the effect of the background emission source on the receptor sites, thereby producing instances of the background class (labelled 0). Thus, an instance measured over hour i, takes the following form:

$$\mathbf{x}_{i,0} \ = \ ^{131}Xe_{i,0},\ ^{133}Xe_{i,0},\ ^{133m}Xe_{i,0},\ ^{135}Xe_{i,0},\ 0. \tag{4}$$

The second phase generates the data for the detonation class (labelled 1). This is done by generating random (in time, space and magnitude) low yield explosions, and measuring their impact on the receptor site. Subsequently, the effect of the detonation is combined with that of the background source over the appropriate

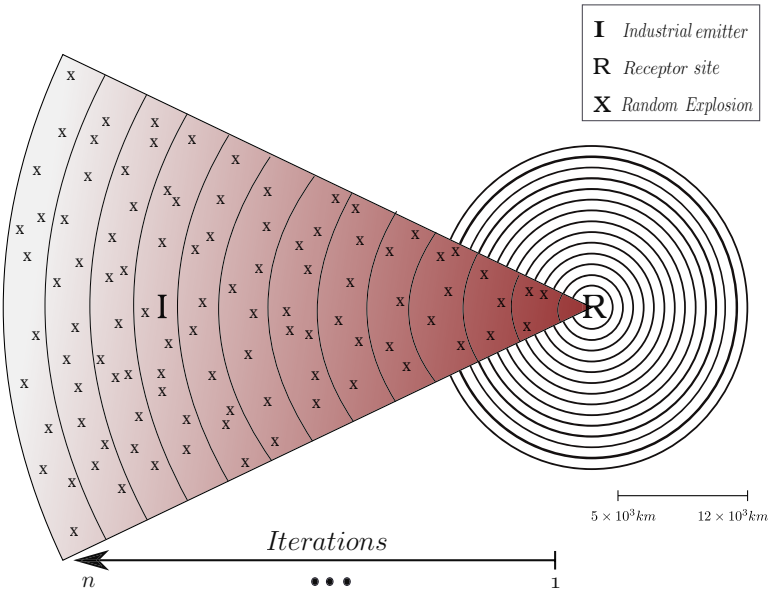

Fig. 4. This figure demonstrates the iterative composition of the simulated domain. In each iteration of the simulation, a fixed number of explosions are probabilistically generated as uniform, random events in time, space and magnitude, and dispersed according to the prevailing meteorology, which may or may not carry the pollutant cloud past the receptor site.

period of time, and written to the dataset with the detonation label. Therefore, a detonation instance measured over hour j, takes the following form:

$$\mathbf{x}_{j,1} = \mathbf{x}_{j,0} \ + \ {}^{131}Xe_{j,1}, {}^{133}Xe_{j,1}, {}^{133m}Xe_{j,1}, {}^{135}Xe_{j,1}, \ 1. \tag{5}$$

3.3 Generated Datasets

A total of 230 datasets were derived and applied to scenario S1 and S2, according to the simulation procedure previously described. More specifically, 10 datasets were generated for each of the 23 detonation ranges, each of which was subsequently divided into training and testing components.

Intuitively, the first scenario presents a slightly easier classification problem, because a set, albeit small, of SE events can be extracted from the application domain and applied to train and/or test the PR systems. More specifically, within this scenario, we assume that the ω_2 class is both identifiable and available in quantities that facilitate the training of binary classifiers. However, in many ways, the classification problem still presents itself as a so-called OC classification task, and thus warrants exploration on both fronts. The datasets specifically contain a 90% background data (ω_1) and 10% explosion data (ω_2).

Alternatively, each set involved in the S2 scenario is divided with 99% background data (ω_1) and 1% explosion data (ω_2). In order to simulate the challenge of manually labelling the instances drawn from class ω_2, and in accordance with the disguised nature of the SE events, all of the ω_2 training instances were erroneously labelled ω_1.

Alternatively, the test sets included appropriately labelled instances from both classes, with proportions following the predefined states-of-nature. This enabled us to assess each classifier's ability to generalize the "real" background data from the noisy training set.

4 PR Solutions

In this section, we present a series of experiments designed to both illustrate the demonstration domain, and to exhibit a first attempt at classifying this sub-category of PR problems.

4.1 Classification Scenarios

As mentioned in the introductory section, within this challenging domain of classification problems, there exist two conceivable scenarios, which we have denoted as S1 and S2. These scenarios explicitly influence the choice of the classification scheme applied to the task of recognizing the SE events.

Intuitively, the first scenario presents a slightly easier classification problem, because a set, albeit small, of SE events can be extracted from the application domain and applied to train and/or test the PR systems. More specifically, within this scenario, we assume that the outlier class is both identifiable and available in quantities that facilitate the training of binary classifiers. However, in many ways, the classification problem still presents itself as a so-called OC classification task, and thus warrants exploration on both fronts.

Alternatively, the second scenario presents itself as a much more difficult PR task, and in many ways more accurately reflects the PR problem suggested by the detection of SE events, in general, and the verification of the CTBT, in particular.

In accordance with the general domain characteristics, as they were originally defined, the data presents itself as a time-series of background measurements that are interwoven with a minute number of SE events. However, unlike the ideal scenario depicted in S1, here we attempt to assume a state-of-nature that is more appropriate for the CTBT task. In particular, we assume that there is a 1% *a priori* probability of a detonation, which, while still an overestimate, is a more accurate depiction, while it still provides insight into the behaviour of PR systems on the class of SE events.

Raising the difficulty further, is the recognition that, in practice, the clandestine nature of the SE events are such that manually identifying a distant clandestine occurrence in the acquired time-series of readings is extremely difficult, if not impossible. Thus, this prohibits the derivation of a labelled training

set, which dictates that practitioners are left to utilize a training set composed largely of background instances, but with a minute number of *unidentifiable* members of the SE event class.

In the absence of a labelled training set, we propose the application of standard OC learners as unsupervised classifiers. When applying OC classifiers to an unlabelled training set, the practitioner must rely on the knowledge of a domain expert to acquire estimates of the *a priori* class probabilities.

In particular, estimates of the state-of-nature are required to appropriately specify the parameters of the OC classifiers, such as the rejection rate, or error rate. This technique aims to prevent the inclusion of the SE event instances in the generalized description of the background class. Our reliance on an error, or rejection rate, presumes that the SE events will reside on the periphery of the background class, and thus, by marginally tightening the generalization of the background class, those instances of the SE event class will no longer be included.

4.2 Classification

Standard PR problems typically assume the existence of data that was drawn independently and identically from the application domain, and that the data can be divided upon class lines into representative sets. The availability of such data facilitates the training of binary classifiers, which have been shown to be proficient at learning class distributions, and thus at labelling novel instances.

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. However, we mention that their implementations were obtained from Weka.

Alternatively, OC classifiers rely on instances drawn from a single class in the derivation of a discriminant function. A broad set of OC classifiers exists in the literature, each of which applies a slightly different strategy to the construction of a binary discriminant function from a single class. However, in simple terms, the process can be articulated as one in which the selected classifier learns to recognize, in some general terms, novel instances that are similar to those viewed during the training process. Thus, novel instances that do not appear to fit into the learned distribution are designated to the ω_2 class.

Although these classifiers were briefly outlined earlier, to summarize:

- The autoassociator (AA), for example, applies a neural network structure to compress/decompress instances of the concept class exclusively. Thus, an unsuccessful compression/decompression results in the instance being assigned to the second class [14].
- Hempstalk *et al.*, in [12], converted the OC classification problem into binary tasks by estimating the distribution of the concept class and generating instances of the non-concept, accordingly. Finally, a standard binary classifier is trained. This process has been denoted the Combined Probability and Density Estimator (PDEN).

- Alternatively, the one-class Nearest Neighbour (ocNN) algorithm [7] learns a *target rejection rate*, τ, where τ is the distance between the two nearest neighbours with the greatest separation in the training data. Subsequently, all novel instances whose nearest neighbours are at greater distances than τ are classified as outliers.
- We have additionally implemented a modified version of the ocNN in Weka, and denoted it as the scaled ocNN (socNN). Contrary to the ocNN, the socNN classifier is capable of learning a model that accounts for the noise in the training set.
- Subsequent research also explored the performance of the often extolled one-class SVM [28]. However, due to the poor results which were generally equivalent to those yielded by the ocNN, it is not included in the present discussion.

4.3 Classifier Assessment Criteria

As discussed in the previous section, this research considers the performance of the classifier within two distinct scenarios. Within each of the scenarios, namely S1 and S2, we considered the performance of the classifier according to a set of criteria. These criteria are discussed in greater detail.

In particular, we examined the general performance of the classifiers across all of the simulated detonation ranges. Performance in this category is particularly important, as, in practice, the detonation ranges are largely unpredictable. The results of this assessment are presented in Sections 5.1 and 6.1. In addition, we explored the performance of the classifier within two shorter detonation ranges, the result of which is presented in Sections 5.2 and 6.2.

The performance of the classifier, as a function of distance, was also examined. The results of this comparison are detailed in Sections 5.3 and 6.3.

Finally, in light of the inherent challenge of distinguishing these two very similar classes according to the four radioxenon isotopes, we were motivated to explore an expanded CTBT feature space. Based on the significant role held by meteorology in affecting the pollutant levels at the receptor site, we surmised that the inclusion of meteorological features would improve the performance of the classifiers. The results of our experiments with an expanded feature space are provided in Sections 5.4 and 6.4.

5 Results: Scenario 1

In this section, we present the results that were obtained according to the four assessment criteria that were motivated in the previous section, on the first classification scenario, S1. We commence our exploration of PR performance by examining the Area Under the ROC Curve (AUC) scores produced by each classifier over the 23 detonation ranges.

5.1 General Performance

In this section, we present a general overview of the performance levels of each of the considered classifiers on the simulated CTBT domain. More specifically, we

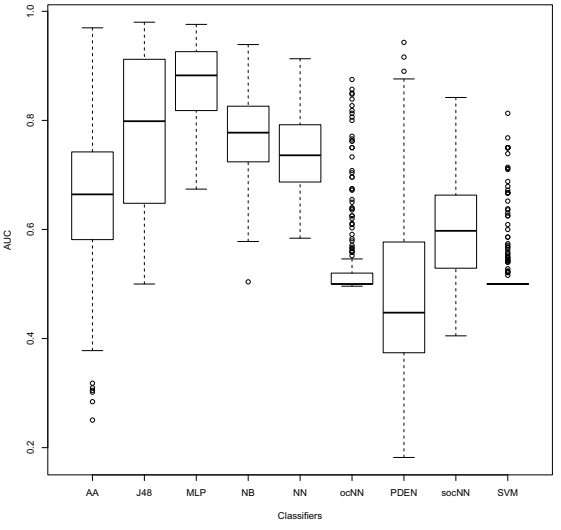

Fig. 5. This figure displays the performance of the nine classifiers, in terms of their AUC scores on the 230 generated CTBT datasets, in the form of a series of boxplots

present an assessment of the five binary classifiers and the four one-class 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 are a particularly insightful overview of the general performance levels.

The results depicted in Figure 5 were compiled as a series of boxplots; one for each classifier.

The solid lines that bisect the boxes represent the median AUC score produced by the particular classifier. The box itself indicates the distribution of the middle half of the AUC scores produced by the classifier. Thus, it stretches from the 25th percentile (at the lower hinge) to the 75th percentile (at the upper hinge). The boxes that are evenly divided indicate that the classifier's scores are evenly distributed throughout the central region. This is, indeed, the case for AA and NB.

The fact that there is no box around the median indicator for the SVM, suggests that nearly all of the AUC results were equivalent, and in this case, approximately 0.5. The relatively large number of circles extending up from the median, individually identify outliers. This suggests that, in general, the SVM classifier performed poorly, but that it occasionally produced anomalously strong results, which stretched slightly beyond 0.8.

Alternatively, the scenario where the median does not produce an even bisection of the box indicates that the distribution of the inter-quartile range is skewed. This is the case, for example, with PDEN, where the upper-quartile is

large, indicating that the points composing the upper-quartile are spread over a larger distance.

The dashed lines, or whiskers, stretch to either the maximum and minimum values, where outliers do not exist, or to 1.5 times the range of the inter-quartile region in scenarios with outliers, such as in the case with the SVM classification results.

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 reiterated 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 being 0.528. Moreover, all four OC classifiers appear to be superior to the SVM when considered in terms of their maximum AUC scores.

When assessing the classifiers according to the boxplot, the median value provides a good indication of their performances, in general. However, most interesting are the ranges of the inter- and outer-quartiles along with the presence of the outliers, when combined with a high median value, as these components provide a strong indication of how likely it is that the classifiers will reproduce the median result.

In these terms, 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, the 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, and does not produce any anomalous results. Indeed, the socNN has a lower standard deviation, and furthermore, its boxplot spans a smaller range.

5.2 Performance on Short- and Long-Range Detonations

In Figure 6, we present the AUC results produced over two detonation ranges of particular interest. The Boxplot on the left in this figure contains the results for the datasets that included detonations ranging from 1,000 km and 5,500 km, while the Boxplot on the right has those with detonations between 5,500 km and 10,000 km. Together, these plots contrast the performance of the individual classifiers in the various detonation ranges. This experimental setup demonstrates one technique through which the performance of various receptor network topologies can be examined. For example, if PR within the second range is found to be a considerable challenge, the shorter range may, perhaps, be considered an upper bound on the acceptable distance between receptors.

There are two factors at play when hypothesizing about classifier performance within these ranges. Intuitively, detonations closer to the receptor site will be more visible at the receptor site, provided the meteorological conditions are such that the emissions are advected in the direction of the receptor. Conversely, detonations that occur farther afield are likely to have a smaller influence on the pollutant levels at the receptor site, leading to a more challenging classification problem. On the surface, then, it appears that nearby detonations should be easier to detect. Indeed, the very near detonations are often easily identifiable. However, the scenario is made more complex by the fact that during the simulation, the industrial

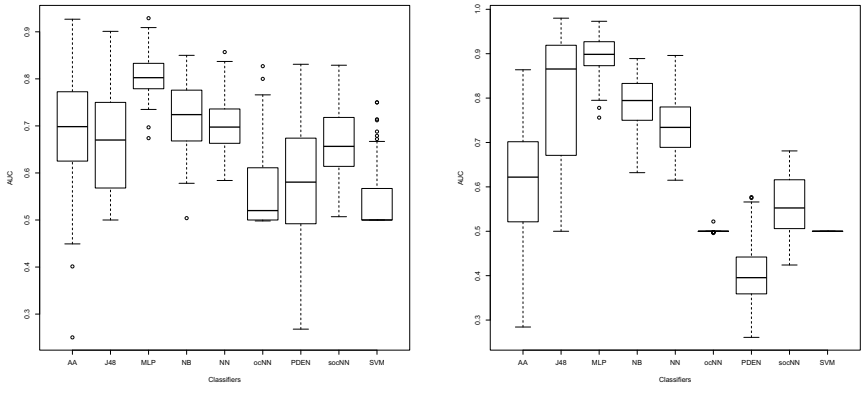

Fig. 6. In this figure, Boxplot (i) displays the performance of the nine classifiers, in terms of their AUC scores for detonations occurring between the distances of 1,000 km and 5,500 km, and Boxplot (ii) displays their performances for detonations between the distances of 5,500 km and 10,000 km

source was positioned approximately in the middle of the shorter range. Thus, there was, in a sense, a great deal of competing background noise to distort the signal.

Indeed, Figure 6 demonstrates that within this scenario it is possible for the performance of the classifiers to improve when detonations occur at greater distances. However, the fact that this only occurred for the binary classifiers, highlights the importance of the second class in the learning process. It turns out that the majority of the binary classifiers are able to, through the training process, utilize the low concentration instances of the detonation class, which resulted from explosions at great distances, to specialize their models to the counter-intuitive point where many of the instances with low concentrations were correctly identified as explosions.

Alternatively, the figure suggests that neither the one-class classifiers, nor the SVM, were able learn a model with this characteristic. Moreover, the SVM exclusively produces AUC scores of 0.5 within the second range, and the ocNN's performance was nearly equivalent. Finally, at greater distances, the PDEN's performance fell even further, with only a minute number of instances exceeding an AUC of 0.5.

Within the shorter range, it is notable that the stronger OC classifiers, namely the AA and socNN, are very comparable with most of the binary classifiers. However, the distinction in favour of the binary learners is emphasized for the larger detonation range.

5.3 Performance as a Function of Distance

In this sub-section, we present the performance of the classifier as a function of distance, where the performance is assessed both according to the AUC and the False Positive Rate (FPR).

A *false positive* occurs when the classifier mislabels a novel instance as a member of the positive class (in this case, a member of the background class), when it is, in fact, a member of the negative class (specifically, a member of the SE event class). Thus, the FPR is the total number of false positives over the total number of negative instances. As a metric, the FPR provides insight into whether the model is overly biased towards the positive class, which is a significant risk when the problem is extremely imbalanced.

These results are particularly interesting, as they provide greater insight into performance trends. Moreover, these suggest a performance scale for successively sparser receptor networks, and enable the interested parties to weigh the cost of receptor stations against the probability of detection.

The performance plots depicted both in Figure 7 and Figure 8 were produced by calculating the ensemble mean of each classifier's performance at the 23 detonation ranges, and then through the extrapolation of a performance function.

Within Figure 7, the MLP classifier is identifiably the superior classifier when compared 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.

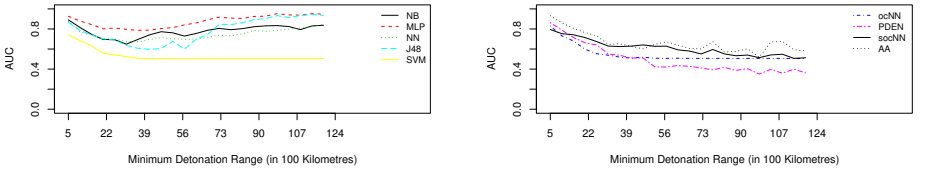

Fig. 7. In this figure, the plot on the left displays the performance of the five binary classifiers, in terms of their AUC scores, as a function of distance. Similarly, the plot on the right displays the performances of the four one-class classifiers as a function of distance, according to their AUC scores.

All of the classifiers, with the SVM appearing as the sole exception, have notable hulls in their performance curves that extend over varying distances and to distinct depths. In each case, a slow descent begins immediately, and is subsequently accompanied by a slow ascent. Alternatively, the SVM classifier suffers from a similar initial decline. However, it fails to recover from the degradation at greater distances.

In each case, the position of the performance hull roughly corresponds to the radial distance between the industrial source of radioxenon and the receptor site. Thus, this suggests that detonations occurring at approximately the same radial distance as that of the primary background emitter are a significant challenge for the detection systems.

The plot on the left in Figure 7 confirms our previous findings, which identified the MLP as the top classifier in this domain, the SVM as the worst, and the remaining three classifiers as contenders for the inner rankings. Indeed, while there are notable differences in the AUC plots for the J48, the NB, and the NN, the fact that their functions cross at numerous points, prohibits the derivation of a general ranking over the entire range of distances.

The plot on the right in Figure 7 presents the performance of the one-class learners as a function of distance. In general, the plot demonstrates that all of the one-class classifiers follow a similar downward trend from their initial peaks, which occurred between 0.8 and 0.9, towards, or beyond in the case of the PDEN, an AUC of 0.5.

Moreover, the performance functions are broadly divisible into two categories. Both the ocNN and the PDEN descend relatively quickly, while the AA and the socNN degrade in a slower, more linear fashion. Therefore, the AA and the socNN are the more suitable of the four one-class learners, with the AA appearing generally superior to the socNN.

The performance of the nine classifiers, measured in terms of the FPR metric, are plotted as a function of distance in Figure 8. In this figure, the plot on the left emphasizes the significant challenge incurred by the binary learners when the detonations occur at a distance similar to the noise source. Although we previously identified the MLP as the strongest binary classifier on this domain, for a relatively broad range (roughly between 25,000 km and 65,000 km), the vast majority of instances, which are truly of the detonation class, were assigned

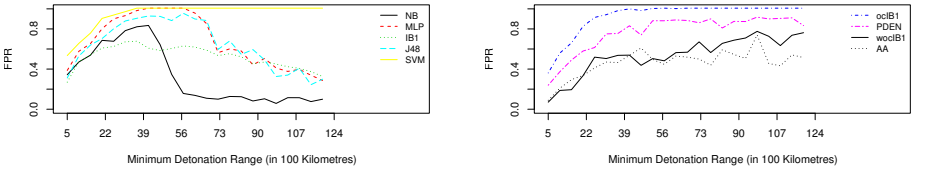

Fig. 8. In this figure, the plot on the left displays the performance of the five binary classifiers, in terms of their FPR scores, as a function of distance. Similarly, the plot on the right displays the performances of the four one-class classifiers as a function of distance, according to their FPR scores.

to the background class. The results are similar for J48. Interestingly, NB has the smallest area under its FPR curve. Thus, it least often identified members of the SE event class as background noise. While we do not consider the FPR results to be individually sufficient for model selection, they do provide some very intriguing insight into the behaviour of the classifiers.

The trends for the one-class classifiers in the plot on the left follow much the same trends previously seen in Figure 7. In particular, the AA and the socNN are superior to the PDEN and the ocNN. However, the distinction between the AA and the socNN is less clear.

5.4 Expanded Feature-Space

Through our exploration of this most interesting of classification problems, we recognized both the inherent challenge presented in the classification of SE events that are interwoven in background noise, and the role of meteorology in effecting the very noise levels that make the task so difficult. Our extensive consideration of this application domain has led us to identify the particularly strong relationship between the wind direction and pollutant levels at the receptor, which suggests a possibly informative feature.

By expanding the standard CTBT feature space to include wind direction, we have produced a significant increase in the AUC. In particular, the top classifiers (MLP, AA, socNN), now demonstrate the ability to detect detonations that, when considered solely on the basis of the four radioxenon measurements, fit into the background distribution with a high probability. This fact is, indeed, depicted for many of the binary and one-class classifiers in Figure 9 and Figure 10.

In particular, while the depth to the hull in the performance of the MLP decreases only slightly, the J48's hull is entirely removed when the wind direction feature is added. Thus, the J48 classification ceases to be affected by the detonation distance when the new feature is included. In addition, its mean AUC is significantly improved.

The NN and SVM classifiers also benefit from the inclusion of the wind direction feature. However, the new feature has a slightly negative effect on the NB. It has been noted in the literature, that many of the PR algorithms, including the MLP, SVM and NB may benefit from normalization of the features [9,32].

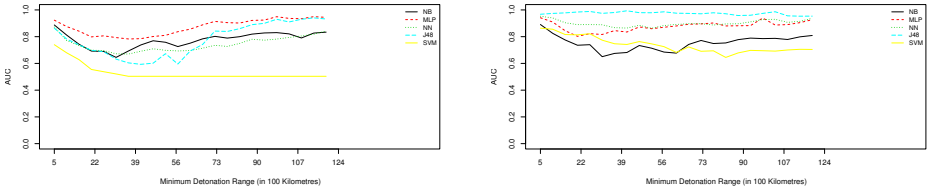

Fig. 9. This figure contrasts the performance of the binary classifiers, in terms of the AUC as a function of distance, on the standard feature-space (see the plot on the left), and when the feature-space is extended to include an assessment of the wind direction (see the plot on the right)

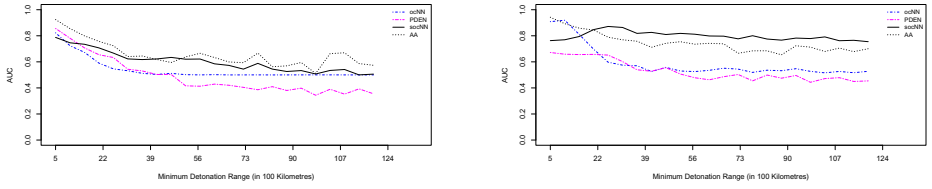

Fig. 10. This figure contrasts the performance of the one-class classifiers, in terms of the AUC as a function of distance, on the standard feature-space (see the plot on the left), and when the feature-space in extended to include an assessment of the wind direction (see the plot on the right)

Thus, it is conceivable that the performance of these classifier may be improved to some degree. However, these results provide a good baseline from which the individual classifiers can be compared.

By expanding the feature-space to include the wind direction, the OC learner, socNN, improves significantly, and becomes, in general, the top learner amongst its peers. The classifier, AA, also improves as a result of the new feature. However, its AUC scores do not increase to the same extent as the socNN.

Similar to the socNN, the PDEN's initial performance is lower in the newly expanded feature-space. However, the majority of its performance function is elevated. Finally, the ocNN benefits the least from the new feature, although, its initial performance is improved.

Thus, in the worst case, the wind direction feature produces marginal improvements in the performance of the four OC learners. However, it significantly improves both the AA and the socNN's ability to perform in scenarios where the detonations occur at distances equivalent to, and beyond the radial distance to the background source.

In Figure 11, a series of boxplots are utilized to facilitate the comparison of classifier performance in the two feature-spaces. Indeed, these results confirm the trends that we have previously identified. Particularly noteworthy is the depiction of J48's performance; this plot emphasizes both the significant increase in the J48's median AUC score, and the impressive stabilization of its classification

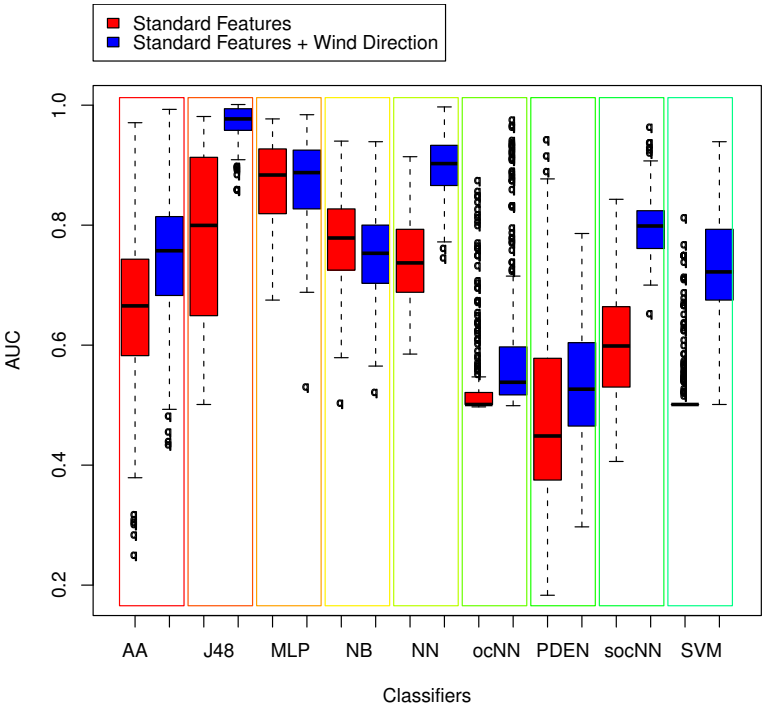

Fig. 11. This figure utilizes a series of boxplots to compare the performance of the nine classifiers and the standard feature-space, and with the extended feature-space, which is augmented by a wind direction indicator

results when the wind direction feature is added. The benefits to the SVM are also well visualized in this figure.

It is, indeed, well demonstrated in Figure 9, Figure 10, and Figure 11 that the additional information has assisted many of the classifiers to overcome the significant challenges inherent in identifying SE events within the field of background noise.

6 Results: Scenario 2

In this section, we present the results that were produced on the four assessment criteria that were motivated, and utilized in the previous sections. In this section, however, we explore the very intriguing classification scenario, which we previously denoted S2. This exploration follows the same structure that was previously applied in the exploration of the first classification scenario. Thus, we begin by examining the AUC scores produced by each of the one-class classifiers over the 23 detonation ranges; we then proceed to consider the performance over the two successive, smaller distances, the performance as a function of distance,

and finally the benefit of expanding the feature-space to include an additional wind direction feature.

6.1 General Performance

In this section, we present a general overview of the performance of the set of one-class classifiers on the simulated CTBT domain. More specifically, we present an assessment of the four one-class classifiers, in terms of their AUC scores on the 230 datasets that covered the 23 detonation ranges.

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

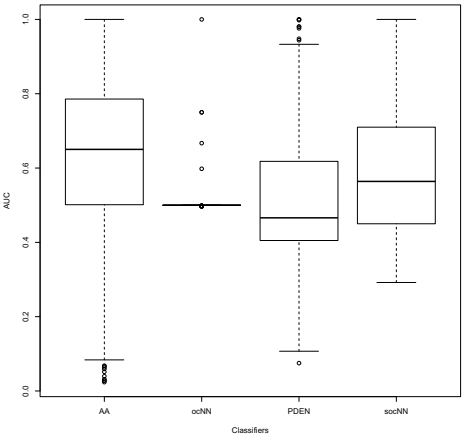

Fig. 12. This figure displays the performance of the four classifiers, in terms of their AUC scores on the 230 generated CTBT datasets, in the form of a series of boxplots

The results that are depicted in Figure 12 were compiled as a series of boxplots; one for each classifier. In addition, 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 both Figure 12 and 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, which appear as outliers in the boxplot. In addition, it is extremely unstable, with results ranging from perfect to near zero.

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

6.2 Performance on Short- and Long-Range Detonations

In Figure 13, we present the results produced over two detonation ranges of particular interest. Specifically, the Boxplot on the left in the figure contains the results for the datasets that include detonations between the distances of 1,000 km and 5,500 km, while the Boxplot on the right has those with detonations between 5,500 km and 10,000 km. Together, these plots demonstrate, contrary to the previous results, that there is little change in performance at greater distances.

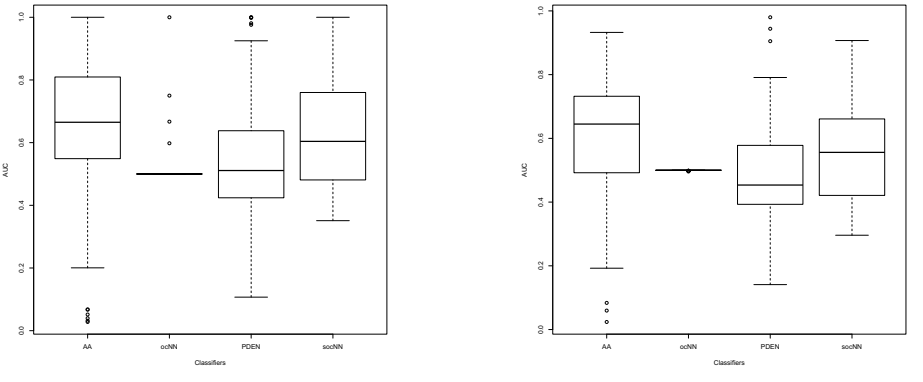

Fig. 13. In this figure, Boxplot on the left displays the performance of the four classifiers, in terms of their AUC scores for detonations occurring between the distances of 1,000 km and 5,500 km, and the Boxplot on the right displays their performances for detonations between the distances of 5,500 km and 10,000 km

6.3 Performance as a Function of Distance

In this sub-section, we present classifier performance as a function of distance. As in the previous section, performance is assessed both according to the AUC and the FPR.

The AA and socNN are, once again, roughly identifiable as the best of the four classifiers in Figure 14 and Figure 15. 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.

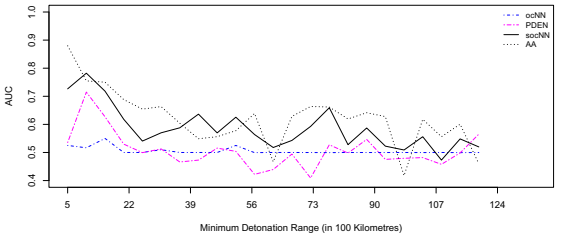

Fig. 14. This figure displays the performance of the four one-class classifiers as a function of distance, according to their AUC scores

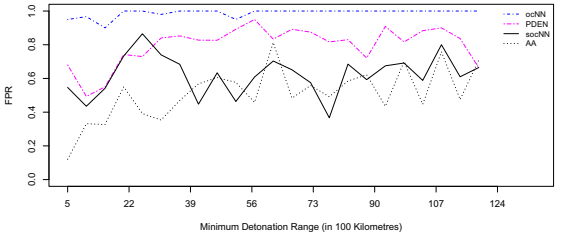

Fig. 15. This figure displays the performance of the four one-class classifiers as a function of distance, according to their FPR scores

6.4 Expanded Feature-Space

In this final section, we consider the benefits of extending the feature space to include a wind direction indicator. In Figure 16, both the original plot of the four classifiers' performances as a function of distance, and their performances on the extended feature-space are plotted. For an alternate view, the comparison is composed of a series of boxplots in Figure 17.

These figures illustrate that both the AA and the socNN significantly benefit from the expanded feature-space. Indeed, the socNN benefits the most, as it becomes superior to the AA for the vast majority of distances, and the variability in its results are significantly dampened.

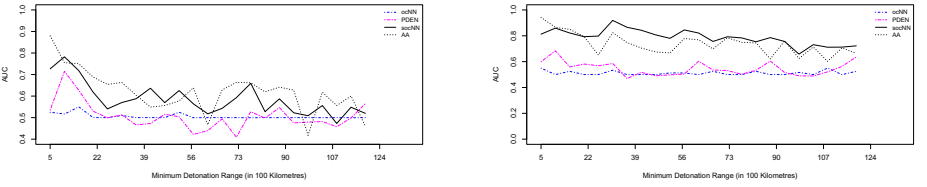

Fig. 16. This figure contrasts the performance of the one-class classifiers, in terms of the AUC as a function of distance, on the standard feature-space (see the plot on the left), and when the feature-space is extended to include an assessment of the wind direction (see the plot on the right)

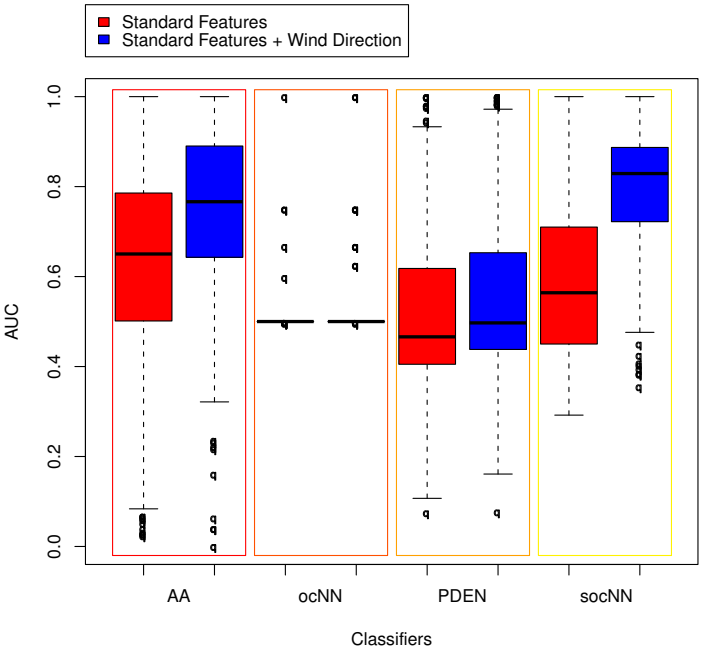

Fig. 17. This figure utilizes a series of boxplots to compare the performance of the four classifiers and the standard feature-space, and with the extended feature-space, which is augmented by a wind direction indicator

7 Discussion

In this section, we consider the results previously reported for the OC classifiers in comparison to those reported for the binary learners. In particular, Section 7.1 compares the two classification strategies within the first scenario, namely S1. Alternatively, the OC classifiers are considered in comparison to the set of standard binary classifiers on scenario S2 in Section 7.2.

7.1 Results: S1

The relatively low mean and median AUC scores produced by the OC classifiers, combined with the considerable variability in their results on the standard CTBT feature-space, particularly in comparison with the top binary learners, clearly illustrate the many challenges inherent in applying OC learning to the derivation of a binary classifier. However, Hempstalk *et al.*, in [12], previously identified similar comparisons between binary and OC learners as "naïve" comparisons, when applied to scenarios that are accurately identifiable as OC problems.

In particular, in so-called OC problems, such as the detection of SE events, the second class is inherently ill-understood due to the fact that a characteristic set cannot be drawn from it. Thus, training and testing a binary learner as if one could draw a representative set from the second class, which is generally assumed when training a binary classifier, provides an upper bound on the classifier's future performance.

The key differences in the performance of the two forms of classifiers is well illustrated in Figures 6 and 7. While the OC classifiers are very competitive on the initial radial ranges, when the detonation occurs further afield, their AUC scores drop considerably in comparison to all of the binary classifiers, with the exception of the SVM. The initial success 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, through the binary learning process they are also capable of drawing on the anomalously low levels, which commonly result from detonations that occurred well beyond the radial distance to the background source, to specialize their decision boundaries such that similar events are recognized as belonging to the SE event class in the future.

The results of expanding the standard CTBT feature-space to include an indicator of the prevailing wind were, in general, very favourable, and lead to improved AUC scores for most of the classifiers, with the NB being the sole exception.

In its essence, the wind direction feature enabled the classifiers to learn the direction of the background source. As a result, the classifiers were able to identify detonations, which occurred at similar radial distances to the receptor site as the background emissions, and thus, had signatures that were similar to the background levels, but were transported from a different direction. This result is identified very clearly in Figure 9, and suggests that the further expansion of the feature-space might additionally improve performance.

7.2 Results: S2

A considerable portion of the previous analysis is applicable to this second, more challenging, classification scenario. Most importantly, the benefits of the extended feature-space were witnessed within S2 as well. However, due to the nature of the problem, only the OC classifiers were applied to this first attempt at performing PR within this new domain.

As a result of the formulation of the problem, we proposed the use of standard OC classifiers as unsupervised learners, and relied on inner mechanisms of the individual classifiers to facilitate the derivation of a model that segregated those instances of the training set that were accurately of the background class from the naïvely/erroneously labelled instances of the outlier class.

It is clear that the instability in performance that is depicted with respect to distance, and which is significantly more apparent in S2 than S1, results both from the erroneous instances in the training sets of S2, and the variability in 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 radial ranges, when considered in terms of the overall means, or medians, the performance of the OC classifiers on the S2 task is only slightly lower than on the S1 task. In addition, this is true if in Figures 7 and 14, we were to conduct our analysis according to a series of best-fit lines.

Finally, as is depicted in Figure 16, in addition to elevating the performance of the top classifiers, the inclusion of the wind direction in the feature-space significantly dampens the variability in their performance. Moreover, Saey, in an extensive study of background radioxenon concentrations in Europe and North America, found that a few outliers representing significant increases in the background concentrations can be expected [26]. These outliers are attributed to alternate background sources, and can be assumed to have arrived at the receptor site via short-lived, and anomalous alterations in meteorology. Based on the standard CTBT feature space, such events, undoubtedly, suggest the detonation of a nuclear weapon. However, provided a sufficient quantity of training data is available, it is conceivable that PR systems functioning with the wind direction feature may appropriately identify outliers of the background class.

8 Conclusions

In this research, we extend the frontiers of novelty detection through 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 one-class 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 not merely impractical - it 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 CTBT. More specifically, we applied the simulation

framework presented in [3], 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 be successfully applied to classify SE events, which are unknown, although present, at the time of training.

Finally, we have added a weighting parameter to the OC nearest neighbour algorithm, thereby significantly increasing its performance on our experimental domain. We have also demonstrated that the expansion of the CTBT feature space significantly improves classifier performance on our simulated data, thus, motivating further exploration of the expansion of the standard CTBT feature space to include meteorological measurements.

References

1. Aha, D.W.: Generalizing from case studies: A case study. In: Proceedings of the Ninth International Conference on Machine Learning, pp. 1–10 (1992)
2. Bellinger, C.: Modelling and Classifying Stochastically Episodic Events. Master's thesis, Carleton University, Ottawa, Ontario (2010)
3. Bellinger, C., Oommen, B.J.: On simulating episodic events against a background of noise-like non-episodic events. In: Proceedings of 42nd Summer Computer Simulation Conference, SCSC 2010, Ottawa, Canada, July 11-14 (2010)
4. Bishop, C.M.: Novelty detection and neural network validation. IEEE Proceedings-Vision Image and Signal Processing 141(4), 217–222 (1994)
5. Bishop, C.M.: Neural networks for pattern recognition. Oxford University Press, Walton Street (1995)
6. Chen, Y., Zhou, X., Huang, T.S.: One-class svm for learning in image retrieval. In: IEEE International Conference on Image Processing, pp. 34–37 (2001)
7. Datta, P.: Characteristic concept representations. Ph.D. thesis, Irvine, CA, USA (1997)
8. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 89(1-2), 31–71 (1997)
9. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, New York (2001)
10. 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)
11. Hanson, S.J., Kegl, J.: PARSNIP: A connectionist network that learns natural language grammar from exposure to natural language sentences. In: Ninth Annual Conference of the Cognitive Science Society, pp. 106–119 (1987)
12. 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)
13. Horton, P., Nakai, K.: Better prediction of protein cellular localization sites with the k nearest neighbors classifier. In: International Conference on Intelligent Systems for Molecular Biology, vol. 5, pp. 147–152 (1997)
14. Japkowicz, N.: Concept-Learning in the Absence of Counter-Examples: An Autoassociation-Based Approach to Classication. Ph.D. thesis, Rutgers University (1999)

15. Kowalczyk, A., Raskutti, B.: One class SVM for yeast regulation prediction. SIGKDD Explorations Newsletter 4(2), 99–100 (2002)
16. 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)
17. Manevitz, L.M., Yousef, M.: One-class svms for document classification. Journal Machine Learning Research 2, 139–154 (2002)
18. Mitchell, T.: Machine learning. McGraw-Hill (1997)
19. Parzen, E.: On estimation of a probability density function and mode. The Annals of Mathematical Statistics 33(3), 1065–1076 (1962), http://www.jstor.org/stable/2237880
20. Petsche, T., Marcantonio, A., Darken, C., Hanson, S.J., Kuhn, G.M., Santoso, I.: A neural network autoassociator for induction motor failure prediction. Advances in Neural Information Processing Systems, 924–930 (1996)
21. Ritter, G., Gallegos, M.T.: Outliers in statistical pattern recognition and an application to automatic chromosome classification. Pattern Recognition Letters 18(6), 525–539 (1997)
22. Roberts, S., Tarassenko, L.: A probabilistic resource allocating network for novelty detection. Neural Computation 6(2), 270–284 (1994), http://www.mitpressjournals.org/doi/abs/10.1162/neco.1994.6.2.270
23. Roberts, S.J.: Novelty detection using extreme value statistics. IEE Proceedings - Vision, Image, and Signal Processing 146(3), 124–129 (1999), http://link.aip.org/link/?IVI/146/124/1
24. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation, pp. 318–362 (1986)
25. 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)
26. Saey, P.R.: The influence of radiopharmaceutical isotope production on the global radioxenon background. Journal of Environmental Radioactivity 100(5), 396–406 (2009), http://www.sciencedirect.com/science/article/B6VB2-4VP1CRK-1/2/ac5135ae3e61e80e9145e24cf1405efd
27. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Computation 13(7), 1443–1471 (2001)
28. Schölkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., Platt, J.: Support vector method for novelty detection. Advances in Neural Information Processing Systems, 12, 582–588 (2000)
29. 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)
30. Tarassenko, L., Hayton, P., Cerneaz, N., Brady, M.: Novelty detection for the identification of masses in mammograms. IEE Conference Publications 1995(CP409), 442–447 (1995), http://link.aip.org/link/abstract/IEECPS/v1995/iCP409/p442/s1
31. Tax, D.M.J.: One-class classification; Concept-learning in the absence of counter-examples. Ph.D. thesis, Technische Universiteit Delft, Netherlands (2001)
32. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann Publishers (2005)
33. Yeung, D., Chow, C.: Parzen-window network intrusion detectors. In: International Conference on Pattern Recognition, vol. 4, p. 40385 (2002)