Clustering Based One-Class Classification for Compliance Verification of the Comprehensive Nuclear-Test-Ban Treaty

Shiven Sharma, Colin Bellinger, and Nathalie Japkowicz

SITE, University of Ottawa
800 King Edward Avenue, Ottawa, Canada
{sshar009,cbell059,nat}@uottawa.ca
http://site.uottawa.ca

Abstract. Monitoring the levels of radioxenon isotopes in the atmosphere has been proposed as a means of verifying the Comprehensive Nuclear-Test-Ban Treaty (CTBT). This translates into a classification problem, whereby the measured concentrations either belong to an explosion class or a background class. Instances drawn from the explosions class are extremely rare, if not non-existent. Therefore, the resulting dataset is extremely imbalanced, and inherently suited for one-class classification. Further exacerbating the problem is the fact that the background distribution can be extremely complex, and thus, modelling it using one-class learning is difficult. In order to improve upon the previous classification results, we investigate the augmentation of one-class learning methods with clustering. The purpose of clustering is to convert a complex distribution into simpler distributions, the clusters, over which more effective models can be built. The resulting model, built from one-class learners trained over the clusters, performs more effectively than a model that is built over the original distribution. This thesis is empirically tested on three different data domains; in particular, a number of artificial datasets, datasets from the UCI repository, and data modelled after the extremely challenging CTBT. The results offer credence to the fact that there is an improvement in performance when clustering is used with one-class classification on complex distributions.

1 Introduction

Compliance verification of the Comprehensive Nuclear-Test-Ban Treaty (CTBT) provides a challenging and an interesting domain for classification. Amongst the technologies used for compliance verification, namely hydro acoustic, infrasound, seismic and radionuclide monitoring [15], the latter provides the only means for unambiguously discriminating a low-yield, clandestine nuclear explosion from other, background events. Thus, in support of the CTBT, monitoring stations with the capability of sampling and measuring the active concentration of four radioxenon isotopes, namely ^{131m}Xe, ^{133}Xe, ^{133m}Xe, and ^{135}Xe by SPALAX technology [14, 4], have been installed at numerous sites across the globe.

The verification challenge lies in discriminating background measurements from those derived from anthropogenic nuclear explosions. The problem is further exacerbated by the fact that measurements from explosions are, in essence, non-existent. Thus, the resulting datasets are, at best, highly imbalanced.

Traditional classification algorithms, which are *discriminatory* in nature since they rely on *discriminating* between all data classes to build models, are known to suffer when presented with imbalance [10]. As a result, one-class (OC) classifiers become more appealing. These methods use data from a single class to build a model, and are based on *recognition*, since they learn to *recognise* data from a particular class, and reject data from all other classes.

For the purposes of compliance verification, OC classifiers aim to learn a description of background data. However, this data comes from a highly complex distribution, and modelling all the various nuances in order to correctly recognise/reject future instances becomes increasingly difficult. This inevitably leads to ineffective performance. A remedy to this problem is to cluster the complex distribution into simpler distributions, and build OC classifiers on these clusters. The resulting combined model should perform more effectively than if we had trained OC classifiers on the original distribution. This idea is illustrated in Figure 1.

In order to examine the effects of clustering for simplifying complex distributions, we conduct experiments using two different OC classifiers, an autoassociator (AA) [8] and a probability density estimator (PDEN) [6], on three different types of datasets: two artificial datasets, seven datasets from the UCI repository, and data modelled after the challenging CTBT domain. The results offer evidence in support of the fact that clustering increases the performance of OC classifiers when dealing with complex distributions.

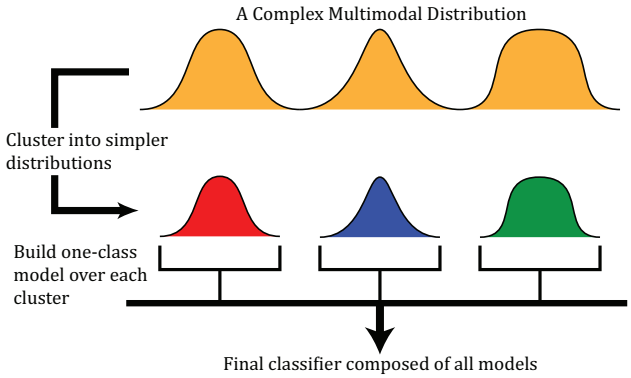


Fig. 1. Framework for One-class classification using Clustering

The remainder of the paper is structured as follows. Section 2 contains an overview of previous work in the field of OC classification, and the use of classification for the verification of the CTBT. A description of the basic framework of the system and how classification is done within it, along with a mathematical formulation and analysis of the framework are presented in Section 3. The artificial dataset, the UCI datasets and the CTBT data are described in detail in Section 4. The experimental framework is described in detail in Section 5. The results of the experiments are presented and discussed in detail in Section 6. Finally, we provide concluding remarks and possible directions for further work in Section 7.

2 Related Work

This section is divided into two parts. In the first part, we provide an overview of OC classification. This is followed by an overview of classification based verification of the CTBT.

2.1 One-Class Classification

Many real-world situations are such that it is only possible to have data from one class, the *target class*; data from other classes, the *outlier classes*, are either very difficult or impossible to obtain. Examples of such domains include those in which there are almost an infinite number of instances from the outlier classes, such as in typist recognition, or those in which obtaining instances from the outlier classes is dependent upon the occurrence of a *rare event*[1], such as the detection of oil spills [9], the inclusion of journal articles for systematic reviews [12], or, as in our particular case, the verification of the CTBT by measuring concentration of radioxenon isotopes. A traditional approach to OC classification is to use density estimation. This is performed by attempting to fit a statistical distribution to the data from a single class (the target data), and using the learnt density function to classify instances as belonging either to the target class (high density values), or to the outlier class (low density values). Parametric approaches rely on reliably estimating the distribution of the data beforehand, a challenging and impractical task given that most real-world data takes a complex distribution. An alternative approach to parametric techniques would be to use non-parametric techniques, such as Parzen Windows. But, as the dimensionality of the data increases, these methods suffer from the well known curse-of-dimensionality problem, whereby the computational complexity for density estimation increases drastically.

There are algorithms designed specifically for OC classification. An example of a OC classifier is the AA, which can be thought of as a compression neural network, where the aim is to try to recreate the input at the output, with the compression taking place at the hidden layers. Hempstalk *et al.*, in [6], describe a

[1] It is likely that the outlier class for classification is the target class in reality. However, we use the term *target class* to denote the *majority class*, and it may or may not be the intuitive target class.

method, PDEN, for estimating the probability density function of a single class by first obtaining a rough estimate of the density of target class, generating an artificial class based on it and then performing binary learning. Yet another example of a OC classifier is the OC Support Vector Machine (OCSVM) [11]. OCSVMs assume the origin in the kernel space to be the second class, and, subsequently, learn a boundary that separates the target class from the origin.

3 Framework: Description and Analysis

The framework consists of two parts: the OC classifier used to model the data, and the clustering algorithm which clusters the data. Training is a two step process; cluster the given data using the clustering method, and then build a model using a OC classifier on each cluster. The final classifier is an ensemble of all the various classifiers built on the clusters. Classification is done by a simple method: If a datum is positively classified by at least one of the models, then it is assigned to the *target* class; otherwise, it is classified as an *outlier*. A mathematical formalization of this framework is presented in the following subsection.

3.1 Mathematical Analysis

Let X represent the set of instances under consideration, and ω be the class to which they belong. What we are interested in obtaining is $P(X|\omega)$, the actual posterior probability density function (pdf). Knowing this can allow for OC classification by imposing a threshold τ on the value of this function, *i.e.*,

$$\text{Classification}(x \in X) = \begin{cases} \text{target}, & \text{if } P(x \in X|\omega) \geq \tau \\ \text{outlier}, & \text{otherwise} \end{cases} \tag{1}$$

However, in practice, the best we can do is obtain an estimate $\hat{P}(X|\omega)$ of $P(X|\omega)$. Given this estimate, and the classifier formulation given in Eq. (1), there are two sources of error that can occur when using $\hat{P}(X|\omega)$:

ϵ_t: The probability that we classify a target instance as an outlier instance (a false negative).
ϵ_o: The probability that we classify an outlier instance as a target instance (a false positive).

Now, we cluster X to obtain c clusters X_i, where $X = \bigcup_{i=1}^{c} X_i$. The clusters may or may not be disjoint. We treat each cluster i as belonging its own unique class ω_i, having its unique pdf $P(X_i|\omega_i)$. Performing OC classification on these clusters is equivalent to obtaining an estimate $\hat{P}(X_i|\omega_i)$ of $P(X_i|\omega_i)$. As before, each $\hat{P}(X_i|\omega_i)$ will have its own two sources of error, namely ϵ_t^i and ϵ_o^i. Let ϵ_t^M and ϵ_o^M denote the error of the combined model, composed of the various models built over the clusters.

Since each cluster represents a simpler distribution as compared to the original distribution, a OC learner should be able to model a cluster more efficiently than the original distribution[2]. In other words, $\forall i \in [1, c], (\epsilon_t^i \leq \epsilon_t) \wedge (\epsilon_o^i \leq \epsilon_o)$, and consequently, $\epsilon_t^M \leq \epsilon_t$ and $\epsilon_o^M \leq \epsilon_o$.

We will now attempt to derive a relationship between the combined model errors, ϵ_t^M and ϵ_o^M, and the error of the single model over X, ϵ_t and ϵ_o, for both cases of error, using the assumption stated in the previous paragraph.

- *Error of False Negatives:* For the combined model, this will occur when a target instance is rejected by all of the cluster models. Since the probability of a single cluster model i rejecting a target instance is ϵ_t^i, and each ϵ_t^i is a mutually independent event, the probability of the combined model rejecting a target instance is $\prod_{i=1}^{c} \epsilon_t^i$. Based on the aforementioned hypothesis, since $\epsilon_t^i \leq \epsilon_t$, we have $\prod_{i=1}^{c} \epsilon_t^i = \epsilon_t^M \leq \epsilon_t$.
- *Error of False Positives:* For the combined model, this will occur when an outlier instance is incorrectly accepted by any one of the cluster models. Since the probability of a single cluster model i accepting an outlier is ϵ_o^i, and each ϵ_o^i is a mutually independent event, the probability of the combined model accepting an outlier instance is $\sum_{i=1}^{c} \epsilon_o^i$. In order for ϵ_o^M to be less than or equal to ϵ_o, in the simplest case, if we assume all ϵ_o^i to be equal, a necessary condition is that each $\epsilon_o^i \leq \frac{\epsilon_o}{c}$. However, given that the distribution of instances represented by the clusters is far simpler than the original, more complex distribution, we assume that all ϵ_o^i will have values to ensure that $\sum_{i=1}^{c} \epsilon_o^i \leq \epsilon_o$.

A theoretical proof of the aforementioned statement will be impossible to obtain, since the error probabilities are dependent on the original distribution, the clusters, the various thresholds and the learning model, all of which are highly variable in practice. Thus, in the subsequent sections, we will conduct a series of experiments in order to obtain evidence that would support the clustering approach.

4 Description of the Data Sets

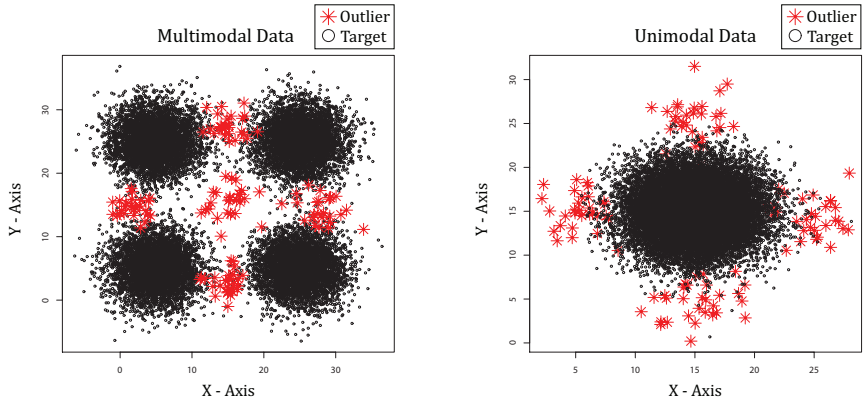
This section provides a description the various data sets used in the experiments. We begin by describing the artificial datasets, followed by the UCI datasets and finally, the CTBT dataset.

Artificial Data: The purpose of using artificial data is to create an idealized data distribution on which we can test the clustering approach to OC classification. By the very fact that it is artificial, we make no attempt to use the results from these datasets to generalize over to practical, real-world problems. However, using artificial data is an important step since it provides a starting point towards more practical, empirical evaluation; it is but a means to an end.

[2] It should be noted that since the cluster models are built only over their corresponding clusters, the distribution of instances that they represent is not the original distribution, but one represented by the corresponding clusters.

There are 20,000 target instances and 125 outlier instances, all part of a bivariate, multimodal distribution consisting of four Gaussian distributions for the target class, and five Gaussian distributions for the outlier class. The high level of imbalance results in the dataset being conducive for OC classification. The standard deviation for both dimensions for the target class is 3. The target class has the following mean vector: $\{(5,5),(25,5),(5,25),(25,25)\}$. The outlier distribution has following mean vector: $\{(15,2.5),(15,15),(15,27.5),(2.5,15),(27.5,15)\}$, and a standard deviation of 2.

Clustering should improve performance in multimodal distributions, but what of unimodal, or simple distributions? In order to investigate this, we use a second artificial dataset. The target class is a unimodal, bivariate Gaussian, having a mean of $(15,15)$, and a standard deviation of 2.75. The outlier class is modelled by four bivariate Gaussians, with $\{(5,15),(25,15),(15,5),(15,25)\}$ as the mean vector. Both distributions are illustrated in Fig. 2.



(i) (ii)

Fig. 2. The two artificial datasets. The larger, denser clusters represent the target classes, whereas the smaller, sparse clusters are the outliers.

UCI Datasets: Although data from the UCI repository does not display the sort of class imbalance that is ideal for OC classification, we have included the following six datasets for completeness; a) diabetes, b) heart disease, c) hepatitis, d) ionosphere, e) thyroid disease, f) sonar, and g) WBCD.

CTBT Dataset: Finally, the CTBT data, which we have presented as our primary domain, is the result of a series of simulated industrial radioxenon emitters and random clandestine tests [2]. The simulation is required here as no "real" clandestine test data exists. The modelling and simulation framework operates in two phases. With respect to the CTBT domain, the initial phase models the affect of industrial sources of radioxenon on the surrounding environment, and accounts for rates of release and variables within atmospheric environment. The second phase models the SE event, in particular, the release of a radioxenon from low-yield, clandestine nuclear test. Alternatively, earthquakes, tsunami waves or unpredicted releases of industrial

pollutants might be modelled as SE events. In this application, the framework models background noise-like non-SE pollutants as Gaussian plumes, and SE contaminants as Gaussian puffs. Both of these Gaussian models have been extensively studied in the literature, and, thus, their strengths and weaknesses are well understood (see [1, 13], for example).

5 Experimental Framework

The experiments are aimed at evaluating the performance of two OC classifiers, AA and PDEN, and their clustered versions[3]. However, we use several binary classifiers (Multilayer Perceptron, the Naïve Bayes classifier, C4.5 Decision Trees, AdaBoost, Bagging and Support Vector Machines) in the experiments conducted on the multimodal artificial dataset, simply to illustrate their performance on highly imbalanced datasets. All classifiers run with their default settings. This is done so as to prevent any bias resulting from the fine tuning the parameters in order to obtain optimal results from specific datasets.

PDEN has also been implemented in WEKA [5], and we use the Gaussian Estimator as the density estimator, and AdaBoost with Decision Stumps as the class probability estimator. Both of these were used with default settings. The binary classifiers have also been used in WEKA.

The experiments with the AA were implemented using the AMORE[4] R package, and run in R[5]. One hidden layer was used for the AA in all the experiments, and the number of training iterations was set to 50. The momentum value was set to 0.99, and the learning rate to 0.01. The number of hidden units for the artificial datasets were set to 4. For all other datasets, they varied from 1 to the number of dimensions of the particular dataset.

The number of clusters for the multimodal artificial dataset was set to 4, given the nature of the dataset. For the unimodal artificial dataset, the number of clusters ranged from 2 to 10. For all other datasets, they varied from 2 to 20.

The performance measure we use is the geometric mean of the per-class accuracies. It is given by $g - mean = \sqrt{acc_1 \times acc_2}$, where acc_i is the accuracy of the classifier on instances belonging to class i. By definition, the metric is immune to class imbalances, and sensitive to the per-class accuracies. It is for these reasons that we selected it for our experiments. Evaluation is done using stratified 10-fold cross validation.

The threshold value for the AA is selected from the set of reconstruction errors over the target training set which maximises the g-mean over the target and outlier training sets. Note that the outlier training set is used only for this purpose; it has no effect on learning.

[3] We present the results graphically and omit the actual g-mean values, as the emphasis is more on the performance trends rather than on the values.

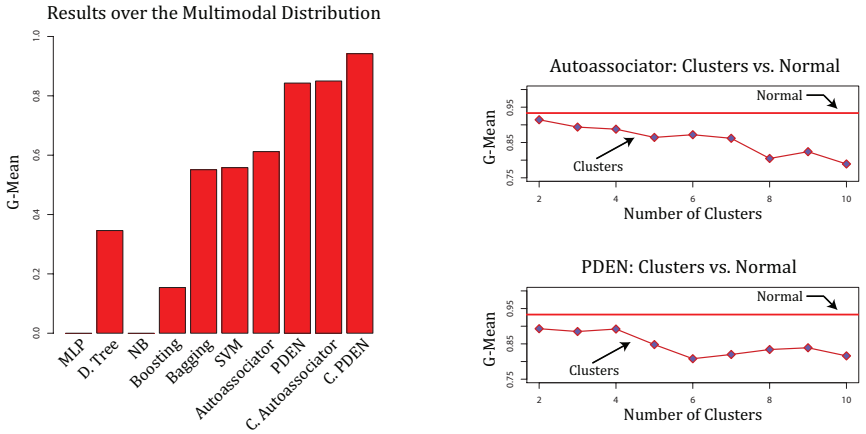
[4] AMORE: A MORE flexible neural network package, `http://cran.r-project.org/web/packages/AMORE/index.html`

[5] The R Project for Statistical Computing, `http://www.r-project.org/`

6 Experimental Results

6.1 Results on Artificial Data:

For the multimodal dataset, the results presented are from running a single AA, a single PDEN, clustered versions of both and a number of binary learning algorithms, and are shown in Figure 3(i). For the unimodal dataset, the results presented are from running only the non-clustered and clustered versions of the OC classifiers[6], and are shown in Figure 3(ii).



(i) (ii)

Fig. 3. The left figure shows the results of various classifiers over the multimodal dataset. The right figure compares the results of the clustered and normal versions of the OC learners over the unimodal dataset.

We stressed earlier that the results over these toy datasets should not be interpreted as a generalization of our method over all domains, and we reiterate that fact here. These datasets merely serve as an illustration of the proposed approach. The multimodal dataset represents a scenario that has a high likelihood of being encountered in practice, i.e., a complex, multimodal target distribution along with a high imbalance ratio between the targets and outliers (in our case, an imbalance ratio of 160:1). The results demonstrate a marked improvement in performance of the OC classifiers when clustering is used, especially in the case of the AA, thus offering evidence in support for the use of clustering. The inherent imbalance of the dataset also demonstrates the failure of the binary classifiers. However, if the distribution is relatively simple, clustering will have a detrimental affect. This can be attributed to the fact that training recognition models on clusters of simple distributions leads to a model that overfits the training data; the simple distribution contains all the information needed to build

[6] Tests on binary classifiers are omitted as the purpose of these tests are only to observe the effect clustering has on simpler distributions.

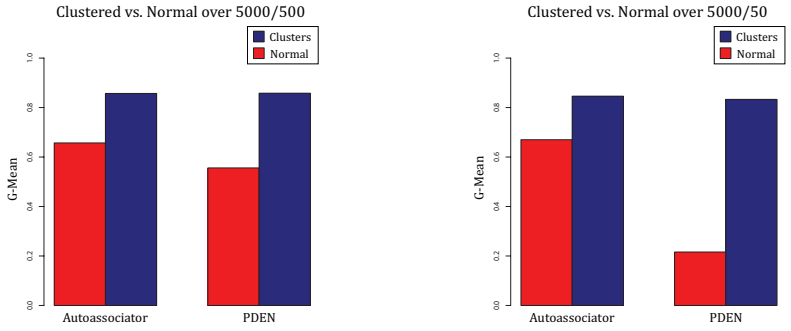
the model, and clustering reduces that, thereby, leading to over-generalisation over sub-regions of the distribution. This in turn causes a higher rate of misclassification of unseen instances.

6.2 Results on UCI Datasets

Although none of these datasets are ideal for OC classification, the target classes have complex distributions, and, as a result, we hypothesize that the *a-priori* clustering will improved the performance of the OC classifiers. Indeed, our experiments confirmed our expectation. In particular, the clustered version of PDEN performs better than the regular version on all datasets, whereas the clustered version of AA performs better on four out of the seven datasets.

6.3 Results on the CTBT Data

We use three CTBT datasets, different only with respect to the level of imbalance between the target (background) and outlier (explosion) classes. The results of the OC classifiers over the dataset with an imbalance ratio of 10 : 1 are presented in Figure 4 (i), over the dataset with an imbalance ratio of 100 : 1 are presented in Figure 4 (ii), and over 250 : 1 are presented in Figure 5.



(i) (ii)

Fig. 4. Results of the clustered and non-clustered (normal) autoassociator and PDEN over the CTBT datasets.

In conducting this research, our domain of primary interest has been that of the CTBT. This is specifically a result of the fact that previous attempts at appropriately applying OC classification methods to the problem have fallen short. We attribute this to the significant degree of complexity present within the background class of the CTBT domain, which inhibits the development of a strong model for recognition-based classification.

With the above in mind, we hypothesized that the utilization of clustering to divide this complex, multimodal distribution into a series of simpler distributions would

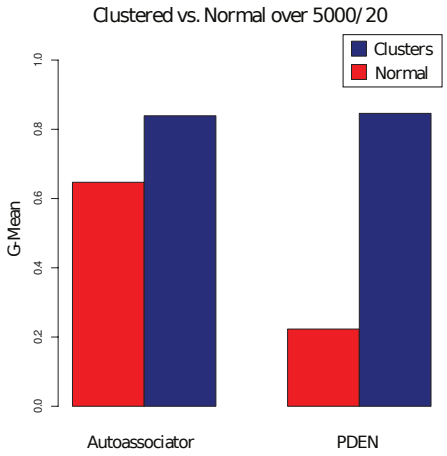


Fig. 5. Results of the clustered and non-clustered (normal) autoassociator and PDEN over the CTBT datasets.

facilitate the development of a superior classification model for the verification of the CTBT. The results presented in the previous section confirm this hypothesis. In particular, both of the explored OC classifiers perform *significantly* better when they are assisted by an *a-priori* clustering phase. It is further encouraging to note that the most significant gains were made on the dataset with an imbalance ratio of $100:1$ and $250:1$, which are both more challenging and realistic. This is particularly apparent in the case of PDEN.

6.4 Statistical Analysis

We perform a statistical test on the results obtained on all the datasets (the multimodal artificial, the UCI datasets and the three CTBT datasets) with AA, PDEN and their clustered versions using a one-sided Wilcoxon Signed Rank Test. This test is more powerful than the paired t-test, as it does not assume normal distributions, assumes commensurability of differences and is less affected by outliers [3, 7]. A significance level of 0.05 was selected. The results are presented in Table 1

Table 1. Results of the Wilcoxon Signed Ranks test for the clustered and regular versions of the OC Classifiers. R^+ represents the minimum sum of ranks, which is taken for those with a positive difference. The minimum sum of ranks should be less than or equal to 10, for $N = 11$.

Classifier	R^+	p-value	α-value	Significant?
Autoassociator	8	0.0122	0.05	**yes**
PDEN	0	0.0004	0.05	**yes**

The results support the fact that for both OC classifiers, the clustered versions outperform the non-clustered classifiers. It is inherently obvious for PDEN, since the clustered version outperforms the non-clustered PDEN over all datasets, thereby giving a minimum sum of ranks as 0.

7 Conclusion and Directions for the Future

Data from many real-world domains come from highly complex distributions, along with high ratios of imbalance between the various classes, presenting ideal scenarios for OC classification. For it to be effective, a OC classifier must be able to model the data from the target class as precisely as possible. In our particular case, we were interested in modelling the background data for the purposes of compliance verification of the CTBT. In order to facilitate this, we investigated the use of clustering for dividing the complex distribution into simpler distributions, which can be modelled more easily by the OC classifiers. We tested the clustering method not just on our own problem domain, but on artificial datasets and datasets from the UCI Repository. The results showed that there is, indeed, an improvement in performance of the OC classifiers, and this was reaffirmed by statistical analysis done using the Wilcoxon Signed Ranks Test.

There are several interesting directions for future research into the use of clustering for OC classification. For the experiments presented here, we used a single clustering algorithm, the k-means algorithm, a simple yet relatively effective algorithm for clustering. With respect to the OC classifiers, we only used AA and PDEN. In future experiments, we will explore the use of clustering algorithms apart from k-means, such as k-medoids or the EM algorithm, different OC classifiers, such as OCSVM and OC nearest-neighbour. It is also likely that, depending on the problem domain, an ensemble of OC classifiers can be used on the clusters, where a different classifier is trained on each cluster.

Research into the detection of anomalous concentrations of radioneuclides in the atmosphere has implications beyond the compliance verification of the CTBT; the recent nuclear crisis in Japan is a stark reminder of the perils of reactor malfunctions. Although this crisis was caused by a natural event and was therefore inherently obvious, in more subtle cases, efficient systems for anomaly detection can act as early warning signlas of impending reactor malfunctions, thereby allowing for timely intervention for rectification and preclusion of large scale, possibly catastrophic, damage.

References

1. Arya, S.P.: Air Pollution Meteorology and Dispersion. Oxford University Press New York, NY (1999)
2. Bellinger, C., Oommen, B.J.: On simulating episodic events against a background of noise-like non-episodic events. In: 42nd Summer Computer Simulation Conference, SCSC 2010, Ottawa, Canada, July 11-14, 2010. Proceedings (2010)
3. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7 (2006)
4. Fontaine, J., Pointurier, F., Blanchard, X., Taffary, T.: Atmospheric xenon radioactive isotope monitoring. Journal of Environmental Radioactivity 72, 129–135 (2004)

5. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter 11(1), 10–18 (2009)
6. Hempstalk, K., Frank, E., Witten, I.H.: One-class classification by combining density and class probability estimation. In: Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science, vol. 5211, pp. 505–519. Springer, Berlin (2008)
7. Japkowicz, N., Shah, M.: Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press (2011)
8. Japkowicz, N.: Supervised versus unsupervised binary-learning by feedforward neural networks. Machine Learning Volume 42, Issue 1/2 42, 97–122 (2001)
9. Kubat, M., Holte, R.C., Matwin, S.: Machine learning for the detection of oil spills in satellite radar images. In: Machine Learning. pp. 195–215 (1998)
10. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: One-sided selection. In: In Proceedings of the Fourteenth International Conference on Machine Learning. pp. 179–186. Morgan Kaufmann (1997)
11. Manevitz, L.M., Yousef, M.: One-class svms for document classification. The Journal of Machine Learning Research 2, 139–154 (2002)
12. Matwin, S., Kouznetsov, A., Inkpen, D., Frunza, O., O'Blenis, P.: A new algorithm for reducing the workload of experts in performing systematic reviews. Journal of the American Medical Informatics Association 17, 446–453 (2010)
13. Simmonds, J.R., Lawson, G., Mayall, A.: A Methodology for Assessing the Radiological Consequences of RoutineReleases of Radionuclides to the Environment. EUR, 1018-5593 ; 15760, European Commission, Directorate-General for Environment, Nuclear Safety and Civil Protection (1995)
14. Stocki, T.J., Japkowicz, N., Ungar, I.K., Hoffman, J., Yi, J.: Summary of the data mining contest for the ieee international conference on data mining. In: Proceedings of the ICDM'08 Data Mining Contest (2008), `http://www.cs.uu.nl/groups/ADA/icdm08cup/booklet.pdf`
15. Sullivan, J.: The comprehensive test ban treaty. Physics Today 51(3), 24–29 (1998)