Expert Systems With Applications 108 (2018) 223–232

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

One-class classification – From theory to practice: A case-study in radioactive threat detection☆

Colin Bellinger[a,*], Shiven Sharma[b], Nathalie Japkowicz[c]

[a] Computing Science, University of Alberta, Edmonton, Canada
[b] Fluent Solutions Inc., Ottawa, Canada
[c] American University, Washington, D.C., USA

ARTICLE INFO

Article history:
Received 29 August 2017
Revised 15 December 2017
Accepted 9 May 2018
Available online 16 May 2018

Keywords:
One-class classification
Imbalanced data
Multiple classifier systems
Small disjuncts
Within-class imbalance

ABSTRACT

Over the years, the acceptance of machine learning as a valuable tool in the real-world has caused much interest in the research community; this is particularly the case as the field of Big Data is coming into prominence. However, real-world data comes with a myriad of challenges, amongst the most prominent of which is the fact that it can exhibit a high level of imbalance. This can come in the form of both within- and between-class imbalance. While a significant amount of research has been devoted to the impact of within-class imbalance over binary classifiers, very little attention has been given to their impact on one-class classifiers, which are typically used in situations of extreme between-class imbalance. During our collaboration with Health Canada into the identification of anomalous gamma-ray spectra, the issue of within-class imbalance in a one-class classification setting was highly significant. In this setting, the imbalance comes from the fact that the background data that we wish to model is composed of two concepts (background no-rain and rain); the rain sub-concept is rare and corresponds to spectra affected by the presence of water in the environment. In this article, we present our work into developing systems for detecting anomalous gamma-rays that are able to handle both the inherent between-class and within-class imbalance present in the domain. We test and validate our system over data provided to us by Health Canada from three sites across Canada. Our results indicated that oversampling the sub-concept improves the performance of the baseline classifiers and multiple classifier system when measured by the geometric mean of the per-class accuracy.

1. Introduction

This paper summarizes and advances our previous research into the one-class classification of gamma-ray spectra. In particular, we explore the difficulties caused by the existence of potentially rare regions in the target class; these regions are more formally known as sub-concepts (Weiss, 2003). They make learning a good model difficult because they may be sparse and separated by areas of low-density or the other class. We consider the benefits of both multiple classifier systems and preprocessoing the data to balance the sub-concepts priors in order to mitigate the negative effects of the less frequent sub-concepts.

The application of this research involved a collaboration with physicists at the Radiation Protection Bureau of Health Canada. The overarching objective is to aid government monitors in predicting the occurrence of a pending radioactive event. Depending on the setting, the specific event may involve a malfunction at a nuclear facility or the nefarious use of nuclear material. Thus, the classification objective is to design and develop a system capable of accurately identifying rare spectra signifying potential dangers.

Our initial attempts with standard one-class classification did not satisfy our objective of high accuracy on the target and outlier classes. By examining the misclassifications using principle component analysis, we were able to identify that the errors were occurring in a common area of the data space that forms a sub-concept associated with heavy rainfall.

To address these false positives, our previous work proposed a multiple classifier system-based approach with a cascade architecture (Sharma, Bellinger, Japkowicz, Berg, & Ungar, 2012). We referred to this as a two-tiered system; the key objective was to simplify the target distribution by separating it into its rain and no-rain sub-concepts. We have subsequently been provided with new

☆ This work was conducted in collaboration with the Radioactive Protection Bureau at Health Canada.

* Corresponding author.
E-mail addresses: cbelling@ualberta.ca (C. Bellinger), ssharma@fluentsolutions.com (S. Sharma), japkowic@american.edu (N. Japkowicz).

https://doi.org/10.1016/j.eswa.2018.05.009
0957-4174/

gamma ray spectra datasets and have identified a limitation in the multiple classifier system approach related to the degree of imbalance in the rain sub-concept. This has inspired us to consider the problem from the perspective of within-class imbalance (Japkowicz & Stephen, 2002), and enabled us to utilize the wealth of research available in the binary class imbalance literature to shape our solution.

The main contributions of this work are to:

- Summarize and extend our previous work on two new radiation monitoring datasets that come from unique domains;
- Explore the problem of gamma-ray spectral classification from the perspective of within-class imbalance, and demonstrate how it can impact the performance of one-class classifiers;
- Propose a solutions to the within-class imbalance problem by extending sampling methods from the binary classification literature to balance the sub-concepts; and
- Show that synthetically oversampling the imbalanced subconcept improves performance more than the multi-classifier approach previously proposed.

2. Related work

Class imbalance has been studied in both extreme cases and moderate cases. The former refers to the situations where no, or almost no, training instances from a concept of interest are available. Alternatively, in the moderate cases, the relative training balance is skewed enough to negatively impact performance. Nonetheless, binary methods with some pre-processing or weighting may still be applied. Both of these veins of research have influenced our solution to the gamma-ray spectra classification problem. As such, we discuss the relevant work below. This commences with the moderate case and proceeds to the extreme setting.

2.1. Binary class imbalance

Class imbalance appears in a wide variety of important and challenging binary classification tasks. Some prominent examples of imbalanced classification problems are: oil spill classification, gene function annotation, and medical and text classification (Akbani, Kwek, & Japkowicz, 2004; Blondel, Seki, & Uehara, 2011; Kubat, Holte, & Matwin, 1998; Nguwi & Cho, 2009). Applications in such areas have demonstrated that it can pose a major challenge for classification systems (He & Garcia, 2009; Japkowicz & Stephen, 2002). In the literature, two forms of imbalance have been identified, namely between-class and within-class imbalance. It has been found that, in many cases, data complexity, such as class overlap, noise and sub-concepts, contribute much of the difficulty to imbalanced problems (Batista, Prati, & Monard, 2004; Denil & Trappenberg, 2010; García, Sánchez, & Mollineda, 2007; Japkowicz, 2001; Prati, Batista, & Monard, 2004).

The issue of sub-concepts is highly relevant in this work. As Stefanowski (2016) highlight, apart from imbalance, the performance of classifiers can be impacted by the presence of sub-concepts (i.e., small dusjuncts). Research in binary classification has established that sub-concepts, particularly rare sub-concepts, can lead to a degradation in classification performance (Japkowicz, 2003). Our work here shows that this is also an issue in one-class classification. For binary classification, Jo and Japkowicz (2004) propose a method for dealing with both within and between class imbalance by clustering and random oversampling. Napierała, Stefanowski, and Wilk (2010), examined further means of managing the affect of noise, overlap and sub-concepts with data cleaning and oversampling based on the local characteristics of the data. They empower rare, but relevant, sub-concepts, whilst removing noise and borderline instances.

The Synthetic Minority Oversampling TEchniques (SMOTE) is the standard method applied for synthetic oversampling in the literature (Chawla, Lazarevic, Hall, & Bowyer, 2003). SMOTE generates new instances of the minority class by interpolating them at random points on the edges connecting nearest neighbors in the minority class. This results in samples created within the convex-hull formed by the minority class. The manifold-based synthetic oversampling sampling method was recently proposed (Bellinger, Drummond, & Japkowicz, 2017); it's approach of modeling data as low-dimensional manifolds is particularly helpful on sparse, high-dimensional domains, as we have here. Samples are generated from the induced manifold, which leads to a better representation of the probability density.

2.2. One class classification

The goal in one-class classification is to induce a binary class predictor, f: $x \rightarrow y$, that learns a functional mapping from the feature vector x to the corresponding label y, where $y \in \{0, 1\}$. Learning takes place on a given set of training examples X sampled from the target class $y = 0$. This is a challenging learning problem because a classifier must be induced, and the model must be selected without seeing examples of the other class $y = 1$.

One-class classifiers typically induce their decision boundaries using one of three modeling paradigms: density-based, recognition-based and boundary-based. Each of these paradigms have been widely applied. Density-based methods have been applied to one-class classification problems, such as diseases and infection detection, and to monitor content (Cohen, Sax, Geissbuhler et al., 2008; Tarassenko, Hayton, Cerneaz, & Brady, 1995; Zuo, Wu, Hu, & Xu, 2008). Reconstruction-base classifiers have been applied to predict failures in helicopter gear boxes, classify documents and to detect nuclear tests (Japkowicz, 1999; Manevitz & Yousef, 2001). One-class support vector machines (SVM) and support vector data description are the standard boundary based methods for one-class classification. These have had a significant amount of success in applications of text, image retrieval and human health (Chen, Zhou, & Huang, 2001; Erfani, Rajasegarar, Karunasekera, & Leckie, 2016; Manevitz & Yousef, 2001; Zhang, Wang, Xu, & Liu, 2006).

Whilst much research has been undertaken to understand the data properties that impact the performance of binary classifiers induced over imbalanced datasets, the relationship between one-class classifiers and data properties has not been as thoroughly considered. A recent study on the impact of data complexity on one-class classifiers is conducted in Bellinger, Sharma, Zaiane, and Japkowicz (2017); the authors highlight that multi-modality resulting from the presence of sub-concepts, as well as class overlap, can cause a significant degradation in the performance of both binary and one-class classifiers as imbalance increases.

In order to make the one-class classifiers more robust to the presence of sub-concepts, Sharma (2016) and Sharma, Bellinger, and Nathalie (2012) demonstrated that by isolating and learning over each sub-concept, better one-class classifier systems can be produced. Isolation is performed by clustering as in Jo and Japkowicz (2004), and a separate one-class classification model is induced for each cluster. This is the motivation for our multi-classifier system, and corresponds to a general approach for building ensembles of one-class classifiers (Jackowski, Krawczyk, & Woniak, 2014; Krawczyk, 2015; Lipka, Stein, & Anderka, 2012); by grouping multiple classifiers into as single system, their collective strengths can be harnessed.

Finally, we note that whilst each of these multiple classifier methods can help to deal with sub-concepts, they implicitly assume that the sub-concepts are well represented. Our results suggest that the relative frequency of the target class sub-concepts have implications on the performance of mutli-classifier systems.

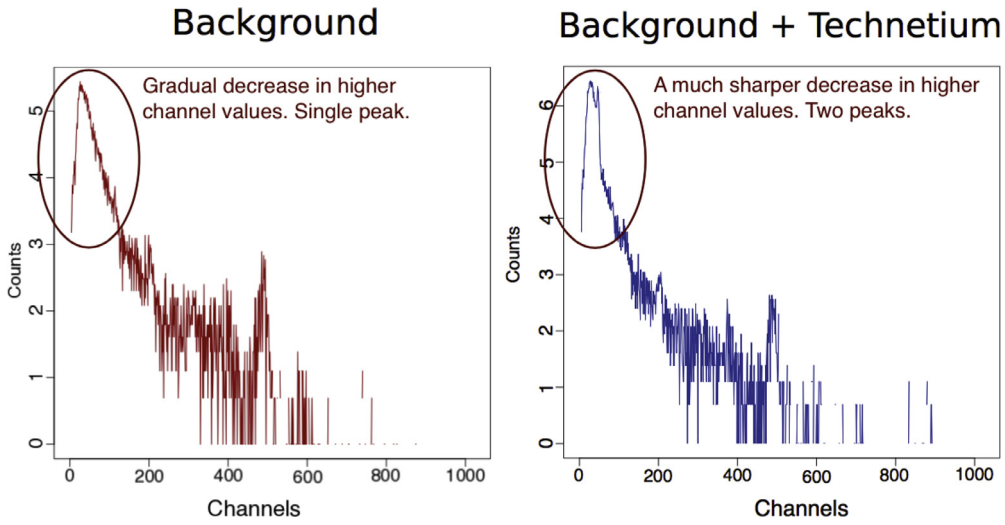

Fig. 1. The plot on the left in log-scale depicts a background instance, and the plot on the right depicts an instance containing the medical isotope Technetium. The key difference in the Technetium signals is in its magnitude and the additional valley and peak after the first peak.

3. Gamma-ray spectra domain

3.1. Background

The physicists at Health Canada were interested in employing a machine learning solution for automatically identifying anomalous gamma-ray spectra as measured by NaI (Sodium Iodide) gamma-ray spectrometers. To this end, they provided us with two types of gamma-ray spectra datasets. The first class of data results from a network of gamma-ray spectroscopes that are intended for environmental monitoring. We refer to this as the national monitoring network. Data from two nodes (Saanich and Thunder Bay) with different radiation backgrounds and complexities are utilized. The second class of data is from the radiation security domain and was collected and monitored during the Vancouver 2010 Winter Olympics. We refer to these as the Saanich, Thunder Bay and Olympics datasets, respectively.

The environmental monitoring network of gamma-ray spectrometers collects data at fifteen minute intervals. The result is a large dataset of gamma-ray spectra each composed 512 channel spectra (each channel can be thought of as a feature in the machine learning context.) The network is designed to detect threats to human health and environment. The vast majority of measurements are solely affected by elements in the local background; these instances are considered to be of no interest. Notably, natural events, such as changes in the wind or heavy rain, can cause significant changes in the background. Alternatively, non-background spectra are cases of interest that should be reviewed by a human analyst. Given the large number of readings resulting from the network each day, machine classification is essential to ensure that the appropriate spectra are given attention in a timely manner.

The Olympics data was recorded at the entrance to Olympic venues in Vancouver. Due to security requirements, this data was collected at one-minute intervals, resulting in significantly more noisy data. Each spectra is composed of 1,024 channels. As no true spectra of interest were recorded during the Games, to evaluate our system, we utilize anomalies that occurred due to the presence of three medical isotopes, namely Iodine, Thallium and Technetium, and a Caesium check-source as the outlier class.

A sample target (left) and outlier (right) instance from the Vancouver dataset is plotted in log form in Fig. 1. In the plots, energy is represented in terms of channels on the x-axis and the counts, which indicate the intensity, are recorded on the y-axis. The subtle difference in the target and outlier can be seen in the shape and magnitude of the lower channels of the spectrum. It is important to note, however, that the shape and location of the change is dependent upon the radioisotope involved in the outlier, along with the quantity and amount of decay that has occurred.

Table 1
Distribution of types of spectra for the three sites. Background and outlier are classes that we want to predict. Background no rain and Background rain is the distribution of the two sub-concepts in the background class.

Type	Vancouver	Saanich	Thunder Bay
Background	39,023	19,063	45,952
Outlier	23	44	731
Background no rain	33,246	18,990	45,820
Background rain	5777	607	905

In addition to the challenge presented by imbalance in the classes, this domain also contains sub-concepts in the target class associated with heavy rain events. The physicists at Health Canada provided us with an additional set of labels indicating whether a given spectra was impacted by rain or not. For simplicity, we refer to these sub-categories as the rain and no-rain sub-concepts. Table 1 reports the distribution of the spectra corresponding to the target and outlier classes, as well as the of the rain and no-rain sub-concepts, for all three data sets. In the Vancouver and Saanich datasets, the class imbalance is too severe for any binary classifier. Whilst there are more outlier examples for the Thunder Bay dataset, it is purely a one-class classification task due to the requirement to detect outliers occurring in all areas of the data-space.

3.2. Challenges of within-class imbalance: the sub-concept of rain

In our work, the gamma-ray spectra impacted by rain had a greater likelihood of being misclassified as outliers than spectra not impacted by rain. More generally, within-class imbalance causes data from the poorly sampled sub-concepts to be treated as outliers by the one-class classifiers. In order to illustrate this phenomenon, let us consider the exemplary domain in Fig. 2. The target class is represented by blue circles and the outlier class is represented by red squares. The target class is composed of a main concept (along the bottom) that is well-sampled, and a sparsely sampled sub-concept spreading upwards in the figure.

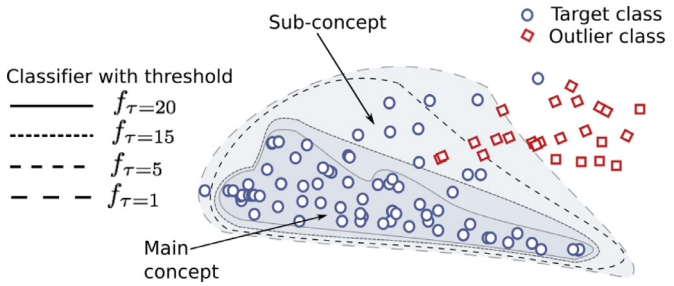

Fig. 2. This figure shows an exemplary one-class classification domain with within-class imbalance. A single classifier is plotted with four different thresholds showing the impact of target and outlier predictions.

Table 2
Accuracy of one-class classifiers over background rain and no-rain spectra.

	Thunder Bay		Saanich		Vancouver	
	Rain	No-rain	Rain	No-rain	Rain	No-rain
AE	0.880	0.903	0.852	0.905	0.775	0.925
ocSVM	0.883	0.901	0.910	0.946	0.804	0.912
MD	0.762	0.801	0.761	0.815	0.669	0.874

Two factors combine to determine the class prediction of a test instance in one-class classification; the rank or likelihood of the instances according to the induced model, and the classification threshold as determined by a rejection rate. If we assume a constant classifier that models the data well, the rejection rate determines the decision boundary. Setting a low rejection rate, τ, *e.g.*, 1%, causes a threshold that keeps the majority (*e.g.*, 99%) of target training instances inside the decision boundary. The risk of a low rejection rate is a higher false negative rate. As the rejection rate is increased, the decision boundary restricts itself to the densest parts of the training set and more target instances fall outside it. Doing so may improve the classification of outliers, and prevent overfitting noise, but can also degrade performance on the target class. This is incrementally depicted for an arbitrary classifier f_τ with rejection rates $\tau = 20$, $\tau = 15$, $\tau = 5$ and $\tau = 1$ in Fig. 2. With respect to the rarer sub-concept, we observe that by increasing the rejection rate, the sub-concept is forced outside the boundary. As a result, classifiers $f_{\tau=15}$ and $f_{\tau=20}$ have increasing numbers of target instances classified as outliers.

Because of the limitation of one-class classifiers, and complexity of our domain, we are required to set relatively high rejection rate to ensure good accuracy on the outlier class. The impact of this is an elevated false positive rate, which contradicts our objective of high accuracy on both classes.

3.3. Demonstration on gamma-ray specta datasets

The visualizations of the gamma-ray spectra datasets using the first two principle components in Fig. 3 demonstrates the existence of sub-concepts in the Saanich and Thunder Bay data. The rain sub-concept is not as clear for the Vancouver data in two-dimensions. It is more visible in three-dimensions, however, we have omited this in the interest of space.

As we demonstrated in Fig. 2, within-class imbalance results in the one-class classifiers treating the data from the minority sub-concept as an outlier, thus causing the accuracy over it to be much lower. We verify this on our gamma-ray spectra domains by examining the accuracy of classifiers on the target rain and no-rain sub-concepts. These results are shown in Table 2. The Mahalanobis distance (MD), reconstruction error of an autoencoder (AE) and a one-class support vector machine (ocSVM) are employed in these experiments to demonstrate their limitations in terms of accuracy on each class. We have selected these three classifiers because they represent the three paradigms, are known to have produced good results in real-world applications, and have been widely applied. They are discussed in greater detail in subsequent section. In each case the accuracy achieved on the target rain sub-concept is lower than the accuracy on the target no-rain sub-concept. We describe the two systems that we developed to address this issue in the following section.

4. Developed systems

In order to solve the one-class classification problem, we tested AE, ocSVM and MD. Together these one-class classification methods cover the spectrum of possible of strategies (recognition-, boundary- and density-based.)

4.1. Applied one-class classifiers

4.1.1. Autoencoder

The autoencoder (AE) is a form of neural network which has an input layer, one or more hidden layers and an output layer (Hanson, 1987). The number of output nodes equals the number of input nodes, whereas the user defines the number of hidden layers and units per layer. The network is trained, layer-wise, in a top-down fashion to optimize an unsupervised objective function using backpropagation with stochastic gradient descent. We use the common objective of minimizing the squared reconstruction error. For more details on designing and training autoencoders, the reader is directed to the textbook of Goodfellow, Bengio, and Courville (2016). To avoid overfitting, denoising is used during training (Vincent, Larochelle, Lajoie, Bengio, & Manzagol, 2010).

Given a query instance x a classification is made using the squared reconstruction error $\left(x - f_\theta(g_{\theta'}(x))\right)^2$. A low reconstruction error indicates that the instance is recognized by the network, whereas a higher reconstruction error indicates that the instance is not recognized and likely comes from a different distribution than the training distribution. Given a fitted model $f_\theta(\cdot)$ and a threshold τ, a query instance x is classified by $\left(f_\theta(x) - x\right)^2 \leq \tau$, where the x is assigned to the target class if the squared reconstruction error is less than the threshold.

4.1.2. One-class SVM

One-Class Support Vector Machines (ocSVMs) (Schölkopf, Platt, Shawe-Taylor, Smola, & Williamson, 2001) are a modification of the standard binary SVMs. Whereas binary SVMs apply an optimization process to find the maximum margin hyperplane that separates the training instances into the two classes, in ocSVM we do not have data from both classes. In this case, the origin of the transformed feature space is treated as the sole instance of the outlier class, and the optimization process finds a hyperplane that maximizes the distance to the origin, subject to a relaxation parameter ν. The free parameter ν behaves similarly to the rejection rate discussed in the context of AE, with the added benefit that it is incorporated into the optimization process so that the decision boundary is determined during training.

The One-Class SVM returns a function f that is positive on instances belonging to S, and negative on those belonging to the complement of S:

$$f(x) = \begin{cases} +1 & \text{if } x \in S \\ -1 & \text{if } x \in \overline{S} \end{cases} \tag{1}$$

In other words, the algorithm for One-Class SVM generates a function f that returns $+1$ in a region S capturing the vast majority of instances (the target class), and returns a -1 in the rest of the region.

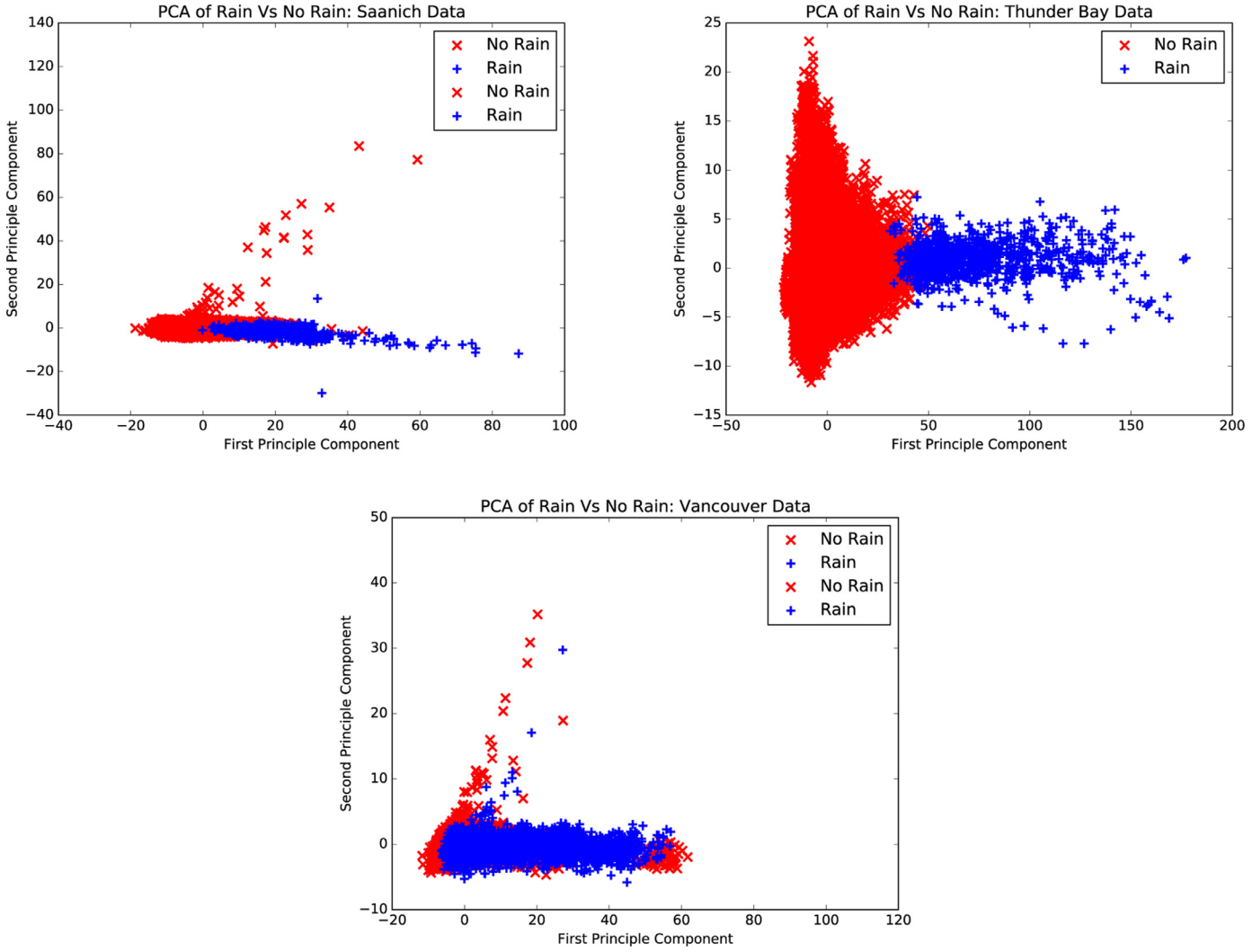

Fig. 3. Presence of the rain sub-concept at Saanich, Thunder Bay and Vancouver datasets.

4.1.3. Mahalanobis distance

The Mahalanobis distance is a parametric distance measure used for Gaussian distributions. By estimating the parameters of the Gaussian distribution for the target class from the training data and setting a threshold on the maximum distance that a query point can be from the sample mean, the Mahalanobis distance forms a simple, but often effective, one-class classifier.

To calculate the Mahalanobis distance we first estimate mean μ and the covariance matrix Σ the target class. With these, the Mahalanobis distance between an instance x from the mean μ is calculated as:

$$MD(x, \mu) = (x - \mu)^T \Sigma^{-1} (x - \mu). \quad (2)$$

Clearly, the larger the Mahalanobis distance between an instance and the mean, the lower the likelihood of that instance being generated by the underlying *pdf*. A Mahalanobis distance classifier (MD) is formed by specifying a distance threshold to the mean of the target class beyond which a test instances is considered as belonging to the outlier class. In line with the previous methodology, the distance threshold τ is selected as the distance which produces the specified rejection rate on a validation set. The rejection rate is set to optimize the true positive rate.

4.2. Addressing the impact of sub-concepts

In the following subsections, we describe two approaches for dealing with the rain sub-concepts in our gamma-ray spectra data. Each of these strategies requires us to have labels indicating which subconcept (rain or no-rain) the training instance belongs to. In gamma-ray spectra classification, this is not an issue because the domain experts can do this in an automated manner using their knowledge of the physical properties of the spectra.

4.2.1. Two tiered multiple classifier system

The first system we developed to address the negative impact of the rain sub-concept in the gamma-ray spectra data has a two-tiered architecture for classification. The motivation for this approach is that it simplifies the target distribution by dividing it into two one-class classification problems. A one-class classifier is induced for each sub-concept in the target class. This is visualized in Fig. 4, where one-class classifier A is induced for the main (no-rain) sub-concept and one-class classifier B is induced for the rain sub-concept. By separating the concepts, we simplify the distribution over which the classifiers are induced.

As described in Sharma et al. (2012), this system utilizes the sub-concept labels (background rain/no-rain) in the first tier of the cascade classifier to induce a discrimination layer for novel spectra. At test time, the binary classifier predicts the sub-concept the

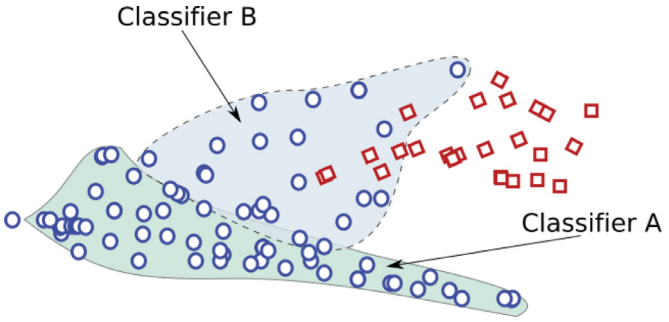

Fig. 4. Managing the subconcept via a two-tiered approach.

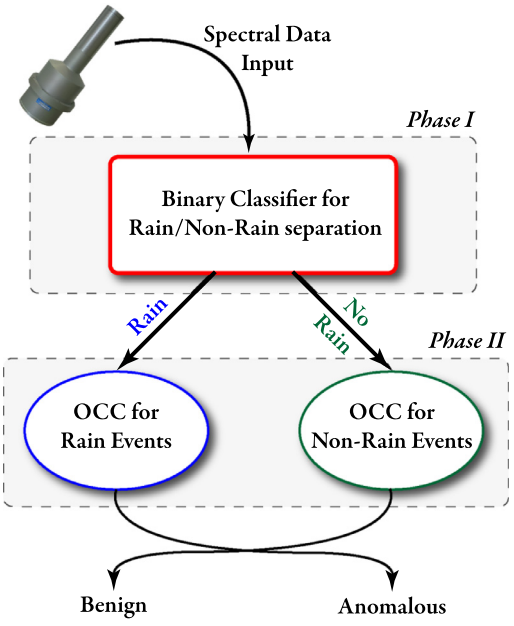

Fig. 5. Training and classification for the two-tiered support.

spectra is associated with, and based on this it is passed to the appropriate one-class classifier. Fig. 5 illustrates the training and classification architecture.

4.2.2. Resampling the small subconcept

Our recent work with the newly shared gamma-ray spectra domains has shown that the multiple classifier approach is negatively impacted by the low frequency of instances in the rain subconcept. Thus, we have proposed an alternative method that utilizes a pre-processing sampling phase to balance the rain and no-rain sub-concepts. This solution is inspired by methods employed to rectify between-class imbalance in binary classification problems, where additional samples are generated in order to produce a better representation of the poorly sampled class. With oversampling, there is always the risk of the newly generated instances overlapping with the other classes. However, in our domain, our analysis of the data properties with the domain experts, along with post-sampling analysis of the data, indicated that overlapping was not an issue.

We apply random oversampling (ROS), which randomly replicates rain spectra, and two methods of synthetic oversampling to generate samples to balance the rain sub-concept. In our context, synthetic oversampling is expected to be beneficial because it generates unique samples to represent the minority sub-concept rather than replicating existing ones or discarding potentially useful instances of the no-rain sub-concept. This is visualized for our exemplary domain in Fig. 6. We perform synthetic oversampling

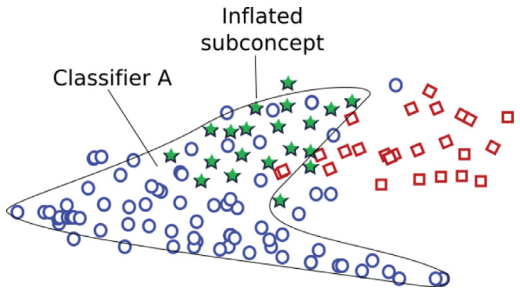

Fig. 6. Managing the subconcept via synthetic oversampling.

with SMOTE and manifold-based synthetic oversmpling (MOS). For completeness, we also apply random undersampling (RUS) to balance the sub-concepts be randomly removing training instances from no-rain sub-concept.

5. Experimental method

The objective of this work is to assess the extent to which our two-tiered multi-classifier approach, and sub-concept empowerment approach, help to improve the overall predictive performance of one-class classifiers by addressing the weakness on the rain subconcept in our radioactive threat detection domain. The remainder of this section describes the methodology applied to test this.

5.1. Datasets

The three gamma-ray spectra datasets from the radioactive threat detection domain described in Section 3 are used in these experiments. As indicated in Table 1, each dataset has a large number of instances and is highly imbalanced. The Vancouver dataset has 1024 dimensions, with 39,023 background instances and 23 instances of the outlier class. The Saanich and Thunder Bay datasets have 512 dimensions. They have 19,112 and and 11,745 background instances, and 44 and 731 outlier instances respectively.

5.2. Algorithms

5.2.1. One-class classifiers

For testing and validating our two systems, we employ three one-class classifiers for modeling the target spectra: AE, ocSVM and the MD. Both the ocSVM and AE require a number of parameters to be set. For ocSVM, we utilized the RBF kernel function. This is the most commonly used kernel function and we found it to be most appropriate for our task. The parameters were optimized in the following ranges: $\nu = [0, 1]$, $C = [2^{-5}, 2^{10}]$, and $\gamma = [2^{-15}, 2^{3}]$ via a random search. The ranges were set based on the recommendations in Chang and Lin (2011). The random search of the parameter space is selected rather than a grid search because it is an efficient method to discover a good parameter set. Nonetheless, we note that the SVM result could be further refined via a grid search. The result would be one of fine tuning, but is not expected to significantly change the results. The model in the random parameter search that produced a rejection rate closest to the objective rejection rate on an independent validation set was kept and applied to the test set.

For the AE, we used bottleneck architecture with three hidden layers of size of size 40, 25, 40 for the Saanich and Olympics datasts, and 75,35,75 for Thunder Bay. In each case, sigmoid, relu, sigmoid activations were used in the hidden layers, and linear activation at the output layer. The network was trained with batches of 250 instances for 3000 epochs with early stopping using the mean square error loss, $l1$ regularization at the hidden layers, and

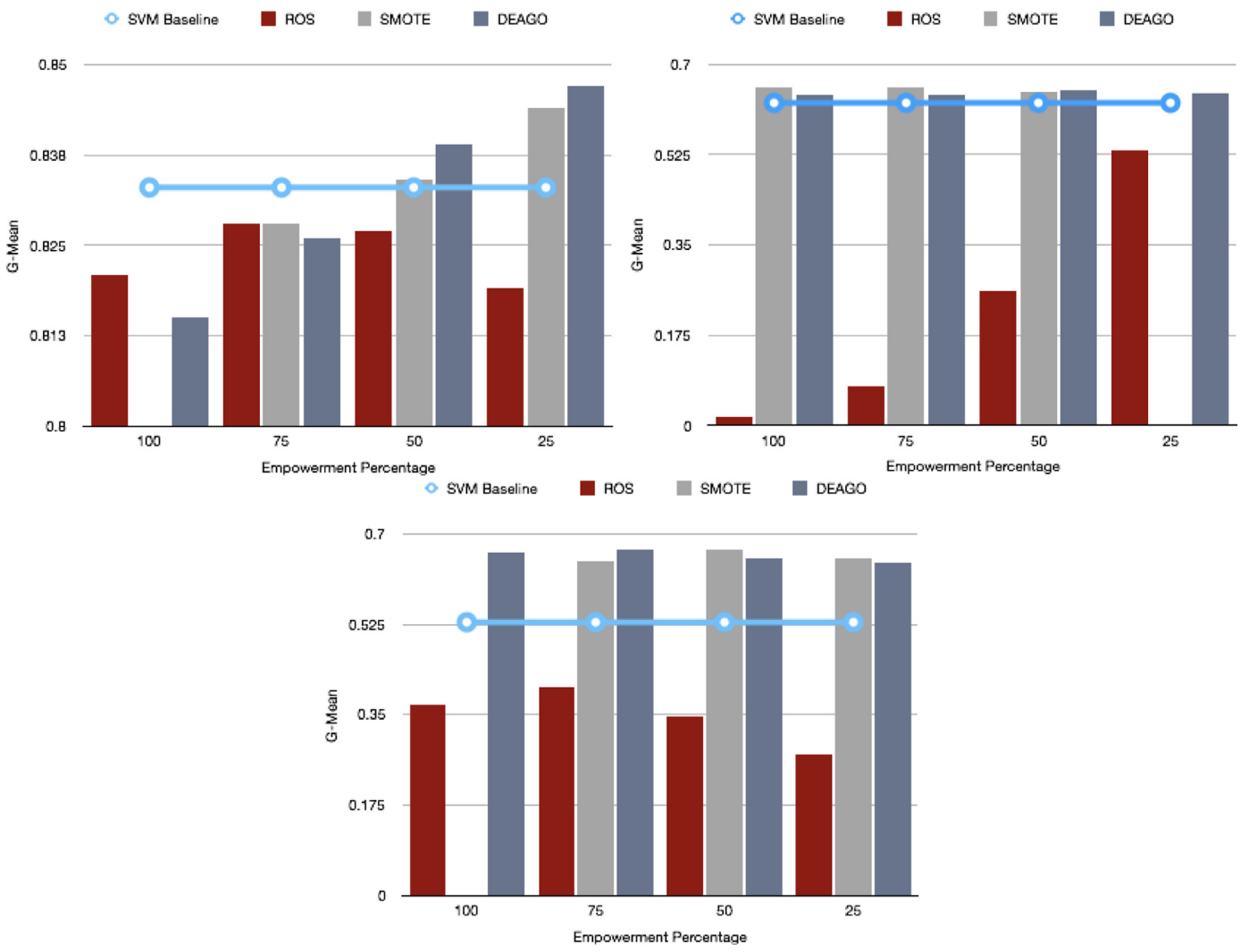

Fig. 7. Baseline SVM performance and the SVM performance after empowerment to 25%, 50% and 75% of the no-rain sub-concept with each oversampling method.

the Gaussian denoising parameter set to 0.01. These parameters were selected based on our previous experience with the data and using autoencoders. The threshold τ was set such that it produced a rejection rate on an independent validation set that matched the user-specified rejection rate.

No parameters were required to be set for the MD classifier. However, for all three classifiers, the threshold τ was set such that it maximizes the per-class accuracy.

5.2.2. Two-tiered system

The two-tiered multi-classifier system includes a binary classifier at the top tier, which predicts whether the test instances belongs to the rain or no-rain sub-concept. In our previous work, we evaluated a full range of binary classifiers for this task and found that the Naïve Bayes classifier produced the greatest improvement in performance of the overall system (Sharma et al., 2012). For this reason, we apply the Naïve Bayes classifier in this work. The three one-class classifier are applied in the second tier of the system according to the process defined above.

5.2.3. Rain sub-concept empowerment system

Sub-concept empowerment experiments applied RUS, ROS, SMOTE and MOS. Each sampling method was applied to alleviate the imbalance by increasing the rain sub-concept to 25%, 50%, 75%, and 100% of the no-rain sub-concept size. In addition, the SMOTE algorithm has a single parameter, nearest neighbours k. We set $k = 7$ based on the proposal in the original work (Chawla, Bowyer, Hall, & Ke., 2002).

The denoising autoencoder implementation of the MOS system includes the parameters previously discussed with respect to the autoencoder one-class classifier. In addition, it includes a step size parameter σ that must be set. This dictates the size of the random step taken along the latent manifold that is used to generate synthetic instances. We set $\sigma = 1$ and performed a random search of the parameter space, keeping the model that minimized the squared reconstruction error as described in Bellinger et al. (2017).

5.3. Evaluation

We use the geometric mean (g-mean) to evaluate the performance of the classification systems. We have selected the g-mean because it provides a combined assessment of accuracy on the target and the outlier class in a single value (Kubat, Matwin et al., 1997). Given the accuracy on the target class a^+ and the accuracy on the outlier class a^-, the g-mean for a classification model f on test set X is calculated as: $g - mean_{f(X)} = \sqrt{a^+ \times a^-}$. This enables us to easily evaluate the extent to which we are achieving the domain objective of producing high accuracy on both classes.

In our results, we report the mean 5×2-fold cross validation of the g-mean. 5×2-fold cross validation is used in place

Table 3
Mean and standard deviation of the 5×2CV G-means for each all systems on each dataset.

	Thunder Bay		Saanich		Vancouver	
	Mean	Std	Mean	Std	Mean	Std
Baseline	0.635	0.098	0.842	0.008	0.873	0.010
TT	0.612	0.022	0.837	0.024	**0.883**	0.005
RUS	0.643	0.010	0.748	0.018	0.731	0.009
ROS	0.402	0.008	0.884	0.018	0.874	0.012
SMOTE	**0.674**	0.025	**0.884**	0.016	0.844	0.020
MOS	0.670	0.032	0.850	0.015	0.876	0.011

of the more common ten-fold version because it has been observed that it has a lower probability of issuing a Type I error (Dietterich, 1998). In addition, k-fold cross validation with larger k values was established with small datasets in mind; the size of the datasets is not a concern here.

In order to perform 2-fold cross validation in a one-class setting, we first divide the dataset into two subsets based on class lines. Therefore, D^+ contains all of the target class instances and D^- contains all of the outlier class instances. Next, D^+ and D^- are each randomly partitioned into two equal sized folds D_1^+ and D_2^+, and D_1^- and D_2^-. In the first round of 2-fold cross validation, the one-class classifier is trained on D_1^+ and tested on the combined set $X = D_2^+ \cup D_2^-$. In the second round, the one-class classifier is trained on D_2^+ and tested on the combined set $X = D_1^+ \cup D_1^-$. This process is randomly repeat five times. Finally, we used the combined 5×2 CV F-test to evaluate the statistical significance of the results (Alpaydm, 1999).

6. Results

6.1. System comparison

The mean and standard deviation of the 5×2-fold CV g-mean for each classification system is shown in Table 3. The first row of this table reports the g-mean produced by the baseline classifier for each dataset. Specifically, on the Thunder Bay dataset, the best baseline g-mean of 0.635 is produced by AE. The best baseline on Saanich is 0.842 and the best baseline on Olympics is 0.873; these are both produced with the Mahalanobis distance (MD). The subsequent rows of the table show the results of the best combination of classifier and corrective strategy (TT, RUS, ROS, SMOTE and MOS). On each dataset, at least one of the corrective measures causes an improvement beyond g-mean of the best baseline classifier. Using the MOS approach to synthetic oversampling led to improvements on all datasets. This shows that it is always beneficial to deal with the rain sub-concept. The best system for each dataset is printed in bold text, and pertain to ocSVM with SMOTE, AE with SMOTE and TT with AE on Thunder Bay, Saanich, and Olympics, respectively.

The Olympics dataset is the easiest dataset with respect to the g-mean. MD produced the highest baseline performance suggesting that the underlying distribution is Gaussian in nature. As shown in Fig. 3, the rain sub-concept does not appear as an overly distinct concept. Furthermore, the relative imbalance between rain and no-rain spectra is less severe as compared to the other datasets, as noted in Table 1. These factors cause all classifiers under all systems to perform relatively well. Due to the stronger baseline performance, only slight improvements can be made to the baseline results. These are produced by TT, ROS and MOS. As we hypothesized, owing to the relatively well represented sub-concepts TT produces the best overall result.

The rain sub-concept in the Saanich dataset is less well represented and more dispersed from the larger no-rain sub-concept. Here, TT does not perform well, and the performance decreases relative to the baseline. On this dataset SMOTE with AE leads to the greatest overall improvement, and each sampling method (ROS, SMOTE, and MOS) produces some improvement.

The Thunder Bay dataset is the most challenging of the three datasets. Once again, the sub-concept on this dataset is very much dispersed from the main rain sub-concept and TT causes a decrease in the g-mean, whereas the synthetic oversampling methods produce noteworthy increases in the baseline from 0.635 to 0.670 and 0.674 with MOS and SMOTE respectively. Interestingly, in this case RUS produces a slight increase over the best baseline, whereas ROS causes a large decrease to 0.402.

To summarize, we hypothesized that the dispersion and complexity in the rain sub-concepts necessitated the use of a resampling-based approach to balance the sub-concepts and improve performance beyond the TT system on the more complex Thunder Bay and Saanich datasets. Our results show that this is indeed the case. We verified the statistical significance of the improvement using the combined 5×2 CV F-test. We were able to reject the null hypothesis which states that there is no difference between the performance of the TT system and corrective resampling with 0.95 confidence on both the Thunder Bay and Saanich datasets.

6.1.1. Effect of sub-concept empowerment

Challenging questions arise with applying oversampling to deal with imbalance. The main question is: how much oversampling is the correct amount? Reinforcing noise and increasing class overlap are risks associated with oversampling, and therefore, it is worth examining the impact of the amount of oversampling on classifier performance.

Our collaboration with domains experts makes us confident that the oversampling methods are not reinforcing noise. However, the other questions are difficult to assess *a priori*. Thus, to evaluate these questions, we plot the changes in performance on each dataset for decreasing levels of empowerment.

These plots, and our previous results, show that synthetic oversampling with SMOTE and MOS is more helpful for SVM than ROS so we focus on them. At each level of empowerment on the Thunder Bay and Saanich datasets, the combinations of SVM and synthetic oversampling cause an improvement beyond the baseline, and the improvement is consistent after empowerment to 25%, 50% and 75% of the no-rain sub-concept.[1] Alternatively, when empowerment is applied to the Olympics dataset, the outcome is very sensitive to the level of oversampling. In particular, empowerment to anything over 50% of the no-rain sub-concept leads to a decrease in the baseline performance. We believe that this results from the fact that the rain sub-concept is relatively well represented in thy Olympics dataset. Thus, excessive empowerment has the negative impact of skewing the underlying distribution.

6.2. Discussion

We considered five means of managing the sub-concept imbalance: a cascade classification approach formed of two tiers, random oversampling, random undersampling, and synthetic oversampling with SMOTE and with MOS. The pre-processing simplifies the system by enabling a single one-class classifier to be trained on a well represented target class. Indeed, our results demonstrate that by pre-processing with sampling we improve upon the baseline classifiers and multiple classifier system when measured by the geometric mean of the per-class accuracy. It should be noted that the presence of noisy samples in the sub-concept could impact classifier performance, as new samples may reinforce noise.

[1] Empowerment by X% implies that the minority sub-concept is X% the size of the majority sub-concept.

In our data, analysis done in with the physicists at Health Canada indicated that there are no noisy samples in the rain spectra, and thus oversampling does not cause any degradation in classifier performance. However, in other domains, this can indeed be an issue, and identifying and dealing with noise would be an essential pre-processing step.

Not all of the combinations of classifiers and corrective measure were equally effective on our data. The choice of the best pre-processing method depends both on the properties of the data and the classifier. Overall, our results show that the sampling approaches are typically superior to the cascade approach on datasets where the sub-concept is a lot rarer. This is particularly the case when ocSVM is selected as the classifier.

Synthetic oversampling is particularly beneficial SVM and AE on our data. In particular, SVM only received a significant benefit from synthetic oversampling, where as AE benefited from all corrective measures. With respect to SVM, this is likely because synthetic oversampling generates new instances that serve as potential support vectors. In the case of MD, replicated samples help to shift the mean to cover more of the rain subconcept. Likewise, by inflating the rain sub-concept through replication seems to be sufficient to assist AE to the corresponding area of the data space.

7. Conclusion

The detection of radionuclides for ensuring public safety and monitoring environmental threats is an important area in the field of gamma-ray spectroscopy. Machine learning has the potential to be a key tool to aid in this task, and we were provided with data from three sites in Canada, Vancouver, Saanich and Thunder Bay, to develop a machine learning solution for detecting anomalous gamma-ray spectra. Our research identified that both between class imbalance (between benign and anomalous spectra), and within-class imbalance (between spectra impacted by rain and those that are not), are major challenges within this domain. While the between-class imbalance necessitates the utilization of one-class classifiers, the within-class imbalance impacts the performance of the utilized one-class classifiers.

Our previous work proposed a multiple classifier system approach to simplify the target distribution, thereby improving performance. In this work, we reevaluate this system on the original dataset and two new gamma-ray spectral datasets. Our latest results show that the multiple classifier system is negatively impacted by the low frequency of training instances in the rain subconcepts in the newer datasets. To this end, we employ concept-empowerment to mitigate the impact of within-class imbalance, by oversampling the rarer sub-concept. The resulting classifier systems improve upon the baseline as well as the multiple classifier system proposed in our previous work.

This work has initiated a discussion on the impact of within class imbalance in one-class classification. However, there is a noticeable dearth of research into understanding the impact of imbalanced sub-concepts in one-class classification. Our future work will examine this topic more directly. In addition, due to our domain focus, we have not fully considered the degree to which multiple classifier systems are, in general, impacted (or not) by within-class imbalance. Given that ensembles and multiple classifiers systems are often found to be very robust, some variation may, in fact, be beneficial for domains impacted by within-class imbalance.

Acknowledgments

The authors would like to acknowledge the funding of the Radiation Protection Bureau at Health Canada, and thank Dr. Kurt Ungar and Rodney Berg for their technical support.

References

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. In *Proceedings of 15th European conference on machine learning* (pp. 39–50).

Alpaydm, E. (1999). Combined 5x2 cv f test for comparing supervised classification learning algorithms. *Neural Computation, 11*(8), 1885–1892.

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter, 6*(1), 20–29.

Bellinger, C., Drummond, C., & Japkowicz, N. (2017). Manifold-based synthetic oversampling with manifold conformance estimation. *Machine Learning*, 1–33.

Bellinger, C., Sharma, S., Zaiane, O., & Japkowicz, N. (2017). Sampling a longer life: Binary versus one-class classification revisited. *Ecml 2017 – 1st international workshop on learning with imbalanced domains: Theory and applications*.

Blondel, M., Seki, K., & Uehara, K. (2011). Tackling class imbalance and data scarcity in literature-based gene function annotation. In *Proceedings of the 34th international ACM sigir conference on research and development in information – SIGIR '11* (pp. 1123–1124). ACM Press.

Chang, C.-C., & Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST), 2*(3), 27.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Ke. , W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). Smoteboost: Improving prediction of the minority class in boosting. In *Proceedings of knowledge discovery in databases: PKDD 2003* (pp. 107–119). Springer.

Chen, Y., Zhou, X. S., & Huang, T. S. (2001). One-class svm for learning in image retrieval. In *Image processing, 2001. proceedings. 2001 international conference on: 1* (pp. 34–37). IEEE.

Cohen, G., Sax, H., & Geissbuhler, A. (2008). Novelty detection using one-class Parzen density estimator. an application to surveillance of nosocomial infections. *Studies in Health Technology and Informatics, 136*, 21–26.

Denil, M., & Trappenberg, T. P. (2010). Overlap versus imbalance.. In *Proceedings of Canadian conference on artificial intelligence* (pp. 220–231). Springer.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10*(7), 1895–1923.

Erfani, S. M., Rajasegarar, S., Karunasekera, S., & Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. *Pattern Recognition, 58*, 121–134.

García, V., Sánchez, J., & Mollineda, R. (2007). An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. *Progress in Pattern Recognition, Image Analysis and Applications*, 397–406.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Hanson, S. J. (1987). Parsnip: A connectionist network that learns natural language grammar from exposure to natural language sentences. In *Proceedings of the ninth annual conference on cognitive science, 1987*.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284.

Jackowski, K., Krawczyk, B., & Woniak, M. (2014). Improved adaptive splitting and selection: The hybrid training method of a classifier based on a feature space partitioning. *International Journal of Neural Systems, 24*(03), 1430007.

Japkowicz, N. (1999). *Concept-learning in the absence of counter-examples: an autoassociation-based approach to classication*. Rutgers University Ph.D. thesis..

Japkowicz, N. (2001). Supervised versus unsupervised binary-learning by feedforward neural networks. *Machine Learning, 42*(1), 97–122.

Japkowicz, N. (2003). Class imbalances: Are we focusing on the right issue? In *Proceedings of the ICML 2003 workshop learning with imbalanced data sets* (pp. 17–23).

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*, 429–450.

Jo, T., & Japkowicz, N. (2004). Class imbalances versus small disjuncts. *SIGKDD Explor. Newsl., 6*(1), 40–49.

Krawczyk, B. (2015). One-class classifier ensemble pruning and weighting with firefly algorithm. *Neurocomputing, 150*, 490–500.

Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine learning, 30*(2–3), 195–215.

Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: One-sided selection. In *ICML: 97* (pp. 179–186). Nashville, USA.

Lipka, N., Stein, B., & Anderka, M. (2012). Cluster-based one-class ensemble for classification problems in information retrieval. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval, SIGIR '12* (pp. 1041–1042). New York, NY, USA: ACM.

Manevitz, L. M., & Yousef, M. (2001). One-class svms for document classification. *Journal of Machine Learning Research, 2*(December), 139–154.

Napierała, K., Stefanowski, J., & Wilk, S. (2010). Learning from imbalanced data in presence of noisy and borderline examples. In *Proceedings of rough sets and current trends in computing: 7th international conference* (pp. 158–167). Springer Berlin Heidelberg.

Nguwi, Y.-Y., & Cho, S.-Y. (2009). Support vector self-organizing learning for imbalanced medical data. In *Proceedings 2009 international joint conference on neural networks* (pp. 2250–2255).

Prati, R., Batista, E., & Monard, M. C. (2004). Class imbalances versus class overlapping : an analysis of a learning system behaviour. In *Proceedings of 3rd Mexican international conference on artificial intelligence* (pp. 312–321). Springer Berlin Heidelberg.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation, 13*(7), 1443–1471.

Sharma, S. (2016). *Learning the sub-conceptual layer: A framework for one-class classification*. Université d'Ottawa/University of Ottawa Ph.D. thesis..

Sharma, S., Bellinger, C., Japkowicz, N., Berg, R., & Ungar, K. (2012). Anomaly detection in gamma ray spectra: A machine learning perspective. In *Proceedings of computational intelligence for security and defence applications (CISDA), 2012 ieee symposium on* (pp. 1–8).

Sharma, S., Bellinger, C., & Nathalie, J. (2012). Clustering based one-class classification for verification of the ctbt. In *Proceedings of 25th canadian conference of artificial intelligence* (pp. 181–193).

Stefanowski, J. (2016). Dealing with data difficulty factors while learning from imbalanced data. In *Challenges in computational statistics and data mining* (pp. 333–363). Springer.

Tarassenko, L., Hayton, P., Cerneaz, N., & Brady, M. (1995). Novelty detection for the identification of masses in mammograms. In *Proceedings of 4th international conference on artificial neural networks* (pp. 442–447). IET.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A. (2010). Stacked denoising autoencoders : Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research, 11*, 3371–3408.

Weiss, G. M. (2003). *The effect of small disjuncts and class distribution on decision tree learning*. Rutgers, The State University of New Jersey Ph.D. thesis..

Zhang, T., Wang, J., Xu, L., & Liu, P. (2006). Fall detection by wearable sensor and one-class svm algorithm. *Intelligent Computing in Signal Processing and Pattern Recognition*, 858–863.

Zuo, H., Wu, O., Hu, W., & Xu, B. (2008). Recognition of blue movies by fusion of audio and video. In *Multimedia and expo, 2008 ieee international conference on* (pp. 37–40). IEEE.