Data-Centric Engineering (2024), 5: e37
doi:10.1017/dce.2024.41

RESEARCH ARTICLE

Handling shift and irregularities in data through sequential ellipsoidal partitioning

Ranjani Niranjan and Sachit Rao

Department of Computer Science, International Institute of Information Technology Bangalore, Bangalore, India
Corresponding author: Ranjani Niranjan; Email: ranjani.niranjan@iiitb.ac.in

Received: 24 April 2024; **Revised:** 29 August 2024; **Accepted:** 17 September 2024

Keywords: dataset shift; dataset skew; ellipsoidal partitioning; small disjuncts; supervised classification

Abstract

Data irregularities, namely small disjuncts, class skew, imbalance, and outliers significantly affect the performance of classifiers. Another challenge posed to classifiers is when new unlabelled data have different characteristics than the training data; this change is termed as a data shift. In this paper, we focus on identifying small disjuncts and dataset shift using the supervised classifier, sequential ellipsoidal partitioning classifier (SEP-C). This method iteratively partitions the dataset into minimum-volume ellipsoids that contain points of the same label, based on the idea of Reduced Convex Hulls. By allowing an ellipsoid that contains points of one label to contain a few points of the other, such small disjuncts may be identified. Similarly, if new points are accommodated only by expanding one or more of the ellipsoids, then shifts in data can be identified. Small disjuncts are distribution-based irregularities that may be considered as being rare but more error-prone than large disjuncts. Eliminating small disjuncts by removal or pruning is seen to affect the learning of the classifier adversely. Dataset shifts have been identified using Bayesian methods, use of confidence scores, and thresholds—these require prior knowledge of the distributions or heuristics. SEP-C is agnostic of the underlying data distributions, uses a single hyperparameter, and as ellipsoidal partitions are generated, well-known statistical tests can be performed to detect shifts in data; it is also applicable as a supervised classifier when the datasets are highly skewed and imbalanced. We demonstrate the performance of SEP-C with UCI, MNIST handwritten digit image, and synthetically generated datasets.

Impact Statement

With classifiers being employed in diverse safety-critical applications, it becomes important to understand the nature of the data on which they are trained, in addition to the underlying algorithm. The problems of identifying small disjuncts, which may be understood as rare and under-represented samples of data, and the occurrence of data shift, where new unseen samples are considerably different than used in training, are considered.

SEP-C is a supervised classifier that partitions a dataset into ellipsoids that mostly contain points of the same label. The rules of classification are determined on the basis of the ellipsoid(s) that contains the test point—this also leads to the calculation of a trust score in the prediction of a label and a possible explanation of why a label was assigned. By suitably changing a single, intuitive, hyperparameter, small disjuncts can be identified. Further, as ellipsoids are in essence Gaussians, well-known methods can be used to detect changes in data.

SEP-C can have an impact in understanding artifacts in data without resorting to heuristics or requiring a priori knowledge. It lends itself naturally to active learning scenarios where a classifier may continuously need to be retrained in the presence of new data.

https://doi.org/10.1017/dce.2024.41 Published online by Cambridge University Press

1. Introduction

The performance of traditional supervised learning algorithms is dependent on the quality of the training dataset. Inherent irregularities such as class imbalance, skewness, small disjuncts, and outliers present in the training data influence the learning and reduce the accuracy of predictions. For example, class-distribution skew can increase the effect of class imbalance, especially around the overlapping region in the dataset. Preprocessing techniques such as exploratory data analysis, Tukey (1977), and data visualization are additionally used to identify such characteristics in the dataset.

The review of data irregularities by Das et al. (2018) observes that traditional classification algorithms assume that each class in the dataset is comprised of one or more subconcepts that are equally represented. This assumption is violated when a small number of rare cases are present in the dataset. These under-represented and rare samples give rise to the problem of *small disjuncts* while learning. While there is no formal definition that identifies a disjunct as being *small*, in the literature, samples of sizes 5, 10, or 15 are termed as *small disjuncts.* Further, errors in prediction due to small disjuncts become pronounced in the presence of noisy data. As also discussed by Das et al. (2018), pruning or reducing the importance of small disjuncts results in eliminating them, which is detrimental to the performance of the model when the rare cases represented by the small disjunct carry valuable learning for the classifier. Alternatively, assigning a new label to the small disjuncts, increases the number of classes in the dataset, converting the problem to multiclass classification. However, the literature does not indicate any practical application of this idea to solve the issues arising out of small disjuncts.

Classifiers for balanced datasets are equally affected by the presence of small disjuncts (Prati et al., 2004). Pruning is ineffective when skewness is observed in the dataset and oversampling to handle class imbalance can result in an increase of small disjuncts. The problem of class imbalance, which is more common and well-studied, has often been attributed to the failing performance of a classifier, but Jo and Japkowicz (2004) observe that cluster-based oversampling improves the accuracy when the issues of class imbalance and small disjuncts are addressed together. Work on the effect of class imbalance, pruning, noise, and training set size in Weiss (2010) highlights that identifying small disjuncts helps to improve the quality of classification to a large extent.

Rule-based algorithms have been used to identify small disjuncts, for example, in inductive systems, a rule consists of several disjuncts, where each disjunct is a conjunctive definition of a subconcept present in the dataset, Holte et al. (1989). The decision-tree classifier performs well on large disjuncts while Genetic Algorithm-based models score better in identifying the small disjuncts. Classifiers built using RIPPER and C4.5 algorithms have been analyzed for the influence of disjunct size and training set size on errors in small disjuncts in Weiss and Hirsh (2000). This work on about 30 datasets, raises an objection to the definition given by Holte et al. (1989) that a small disjunct is one that correctly classifies a few training examples. It emphasizes the need for a threshold for size that is also related to error rate, which can correctly define small disjuncts. Non-rule-based classification algorithms, for example, support vector machines (SVM) and k-nearest neighbor (kNN) classifiers, fail to identify these small disjuncts.

In this paper, small disjuncts are identified using a classifier proposed by the authors, termed as the sequential ellipsoidal partitioning classifier (SEP-C; Niranjan and Rao, 2023a). Given a dataset with points of two labels, SEP-C uses convex methods to find several hyperplanes iteratively, unlike a single separating hyperplane in traditional SVMs or its variants, so that the points of different labels lie on either side of each hyperplane. In each iteration, *nonoverlapping* minimum volume ellipsoids (MVEs; Boyd and Vandenberghe, 2004; Sun and Freund, 2004; Kong and Zhu, 2007) are found using the reduced convex hull (RCH) algorithm (Bennett and Bredensteiner, 2000) to cover the points on either side of the hyperplane. After each iteration, the covered samples from the dataset are removed and the process is repeated until all the data points in the training set are exhausted or the independence-dimension (I-D) inequality is violated (Boyd and Vandenberghe, 2018). The removal of points in each iteration renders this approach as a sequential one.

SEP-C allows for a user-defined number of points of one label to be contained in the MVE of points of another label; this becomes essential when a dataset is heavily overlapped. This number, denoted by n_{Imp}, is the only hyperparameter used in SEP-C. With the introduction of such “impurities,” the dataset can be partitioned finely. In SEP-C, the presence of such impurities is interpreted as being small disjuncts, as these form a subset of points belonging to a label that have properties that are different than the majority of points of that label, hence, rare and under-represented. Thus, by choosing $n_{\mathrm{Imp}} \geq 10$ (the accepted size of a small disjunct), it is possible to find small disjuncts of sizes $\leq n_{\mathrm{Imp}}$; we show that such small disjuncts exist in some of the publicly available datasets. Unlike rule-based methods, where, a small training set is used to train the classifier to find small disjuncts in the much larger testing set, in SEP-C, the entire dataset is used to identify them. As will be discussed, the proposed method is not dependent on the underlying data distribution and hence is immune to the presence of skewness in the training data.

In addition to identifying small disjuncts, we apply SEP-C to detect shifts in data distributions, for example, a covariate shift which may be caused by changes in external conditions, sensor failures, or changes in the environment; such changes may be expected in applications such as autonomous cars and surgical robots and can occur in unanticipated ways or gradually. Such shifts lead to samples that are generally termed as being out-of-distribution (OOD) and as they affect the entire dataset, it can be considered to be different than outliers in the training set. Hence, the problem of data shifts is different from handling outliers. Detecting an onset of data shift helps to identify when the model requires retraining without having to endure poor performance. Guerin et al. (2023) term a data point as OOD if its label does not belong to the predefined labels handled by the classifier. While they identify the need for a threshold to detect OOD samples, finding such a threshold itself is challenging. Overemphasizing the detection of OOD can result in the rejection of correct predictions from the classifier. The work highlights the importance of separating the OOD from the in-distribution training samples which makes detection of OOD data a crucial step in training a classifier.

State-of-the-art classifiers assume that new data samples follow the same distribution as the training data. However, when a shift occurs, there is naturally a decline in the performance of the classifier resulting in unreliable predictions. Deep neural networks (DNNs) that use softmax layers are known to provide overconfident wrong predictions for the OOD data. Lee et al. (2018) describe methods to detect OOD samples and propose a technique that can be applied to any pretrained softmax neural classifier and for in-class incremental learning. This method defines a confidence score that is based on the Mahalanobis distance of the sample from the closest class-conditional distribution. This score controls the high confidence from the softmax layer by reducing the score if the sample is situated far from the class, thereby identifying it as potentially an OOD sample. This method has also been used for adversarial sample detection. Recent research in active learning (AL) promotes a human intervention to take action when the learning model encounters an anomaly. In such applications, Barrows et al. (2021) propose a data distribution shift-detecting method that triggers AL when the model detects a deviation in the arriving data. In this method, the confidence scores from parallel softmax layers are added and a threshold is fixed to indicate the known zone of operation for the classifier. The AL is triggered when the samples produce a confidence score when the threshold is violated.

In another attempt to discover OOD samples, Sastry and Oore (2020) identify inconsistencies between activity patterns and predicted class and use Gram matrix values to achieve superior OOD detection. The Gram matrix values are compared against the range observed for the training data and used along with softmax output to achieve the desired detection; standard metrics, such as true negative and positive rates, accuracy, and AUROC metrics are employed to evaluate the results. Li et al. (2021) employ Bayesian online learning to detect shifts. In this method, a binary change variable is used for the informative prior so that any shift in the distribution is recorded. The authors assume that the samples are independent and identically distributed (i.i.d.), which may not be always true in reality and no assumption is made about how frequently the shifts occur in the test samples. The model is updated to the new distribution when change is detected and erases the past information.

In SEP-C, dataset shift is detected by the expansion of the MVE that is closest to the new data point, if it occurs. According to the rules of classification stated in Niranjan and Rao (2023a), if a new data point does not lie in any of the existing MVE partitions, the MVE closest to it is expanded to cover it and the data point is given the label of those carried by the majority of the points in that MVE. Note that SEP-C finds

https://doi.org/10.1017/dce.2024.41 Published online by Cambridge University Press

nonoverlapping MVEs containing points of the same label and at most n_{Imp} of the other. Thus, a dataset shift can be detected if all new samples continue to be covered by the expanded MVE and not in any of the original MVEs. Indeed, the question that should be asked is: “Is there a threshold for MVE expansion to detect the shift?” In this work, this question is answered by resorting to the univariate Kolmogorov–Smirnov (KS) test to detect changes in distribution. The application of this test is justified for two reasons: i. the MVEs are Gaussians and hence, the local distribution of data is known; and ii. by transforming the data along the eigenvectors of the MVE, they are transformed to n univariate Gaussians, when the dimension of the data is n. It is remarked that the Mahalanobis measure and other KS test variants can also be employed once the MVEs are known. The authors of Rabanser et al. (2019) attempt to predict OOD data and expect a fall in accuracy of the classifier in its prediction. When the ellipsoidal partitions found by SEP-C are used for classification, if the ellipsoids are not expanded, a similar drop in classification accuracy is observed. A marked change in the ellipsoidal parameters is used to identify the OOD nature of the incoming unseen data samples.

The main contributions of the paper are as follows:

1. use of convex methods to partition a training dataset into multiple ellipsoids that contain mostly points of the same label,
2. identify under-represented points of one label that are “different” than most points of that same label —the small disjunct problem, and
3. detect shift in the data based on the expansion of the ellipsoids.

As discussed in Niranjan and Rao (2023a) and will also be shown here, SEP-C is immune to dataset irregularities such as skewness and imbalance; further, the underlying distribution or the new one (in case of shift), is also not required for classification. The results of this paper, mainly the detection of dataset shift, are an extension of those presented by the authors in Niranjan and Rao (2023b).

The paper is organized as follows: in Section 2, SEP-C is introduced. In Sections 3 and 4, the properties of SEP-C being independent of skew and its ability to identify small disjuncts are discussed, respectively; results on identifying small disjuncts in publicly available datasets are presented in Section 4.1. Dataset shift detection is demonstrated using SEP-C in Section 5 for a 2D synthetic dataset as well for the MNIST dataset, where the images are distorted in three different ways. Concluding remarks are provided in Section 6.

2. Methodology

2.1. Preliminaries

Ellipsoidal approximation for datasets is widely used for clustering and outlier detection. Ellipsoids provide an advantage of being regions that are bounded in the feature space, unlike the leaves of a Decision Tree, which can be unbounded; this boundedness, and a resulting change, forms the basis for identifying dataset irregularities, as shown in this paper. All ellipsoid-based problems can be posed as convex optimization problems, which by their nature, have a unique global minimum; moreover, the solution is guaranteed to converge, if found to be feasible. The work of the authors, Niranjan and Rao (2023a), contains a detailed literature survey and the benefits offered by adopting an ellipsoidal partitioning approach for classification, using SEP-C.

The SEP-C algorithm has the advantage of a single hyperparameter usage, unlike many of the state-of-the-art techniques, and implicitly contains multiple hyperplanes which make possible the classification of nonlinearly separable datasets without the need for the “kernel trick.” SEP-C provides a trust-scored based prediction which enhances the explainability of the classifier and reveals dataset nuances such as distribution skews, small disjuncts, and OOD data. Further details on the applications of SEP-C in trustworthy predictions can be found in Niranjan and Rao (2023a).

https://doi.org/10.1017/dce.2024.41 Published online by Cambridge University Press

2.2. Sequential ellipsoidal partitioning classifier

Consider the dataset in Figure 1a, with points denoted by the sets $\mathcal{X} = \{x_i\}, \ \ x_i \in \Re^2, \ \ i = 1, \cdots, N$ and $\mathcal{Y} = \{y_j\}, \ \ y_j \in \Re^2, \ \ j = 1, \cdots, M$, where $N \geq M > 2$. The points in set $\mathcal{X}$ have label L_{+1} and those in $\mathcal{Y}$ have label L_{-1}. The MVEs, denoted by $\mathcal{E}_X$ and $\mathcal{E}_Y$, respectively, can be found such that each ellipsoid contains points of the respective set either in its interior or its boundary. By expressing an ellipsoid in the form $\mathcal{E} = \{z \mid \|\mathbf{A}z + \mathbf{b}\| \leq 1\}, \ \ z \in \Re^2$, for example, for set $\mathcal{X}$, $\mathcal{E}_X$ is found by solving the convex optimization problem (CP)

$$\begin{array}{ll} \min & \log\det\left(\mathbf{A}^{-1}\right) \\ \text{subject to} & \|\mathbf{A}x_i + \mathbf{b}\| \leq 1, \quad i = 1, \cdots, N, \end{array} \tag{1}$$

with the symmetric positive definite (SPD) matrix $\mathbf{A}$ and the vector $\mathbf{b}$ as the variables (Boyd and Vandenberghe, 2004). It is highlighted that for the CP (1) to be feasible, the I-D inequality has to be satisfied, that is, $N > 2$ (Boyd and Vandenberghe, 2018).

As can be seen in Figure 1a, the ellipsoids $\mathcal{E}_X$ and $\mathcal{E}_Y$ intersect each other. SEP-C now partitions the dataset such that the MVEs for each dataset become nonintersecting by applying the RCH algorithm (Bennett and Bredensteiner, 2000). First, the matrices $\mathbf{X} \in \Re^{N \times 2}$ and $\mathbf{Y} \in \Re^{M \times 2}$ that contain the datapoints x_i and y_j, respectively, are defined. Next, the RCHs of the sets $\mathcal{X}$ and $\mathcal{Y}$ are found. These are the set of all convex combinations $\mathbf{c} = \mathbf{X}^T\mathbf{u}$ and $\mathbf{d} = \mathbf{Y}^T\mathbf{v}$, respectively, where $\mathbf{u} = [u_i] \in \Re^2, \ \ \mathbf{v} = [v_i] \in \Re^2$; $\sum u_i = 1, \ \ 0 \leq u_i \leq D, \ \sum v_i = 1, \ \ 0 \leq v_i \leq D$; and the scalar $D < 1$, which is a design parameter. The RCH algorithm finds the closest points in each RC-Hull by solving the CP

$$\begin{array}{ll} \min\limits_{\mathbf{u},\mathbf{v}} & \frac{1}{2}\left\|\mathbf{X}^T\mathbf{u} - \mathbf{Y}^T\mathbf{v}\right\|^2 \\ \text{subject to} & \mathbf{e}^T\mathbf{u} = 1, \ \ \mathbf{e}^T\mathbf{v} = 1, \ \ 0 \leq \mathbf{u}, \ \ \mathbf{v} \leq D\mathbf{e}. \end{array} \tag{2}$$

The vector $\mathbf{e}^T = [1\,1\cdots]$. Now, if the solution to this CP exists for some $D < 1$, the RC-Hulls do not intersect and thus, the line normal to the line connecting the closest points is the separating hyperplane. Solving the CP (2) results in the nonintersecting RCHs for the two sets, as shown in Figure 1b.

In SEP-C, the RCH algorithm is implemented differently. Beginning with $K = \min(N, M)$ and $D = (1/K)$, the CP (2) is solved iteratively, where K is reduced in each iteration until such time that the RCHs of both sets do not intersect or the RCH of points of one label contain at most n_{Imp} points of the other. To determine that the RCHs indeed do not intersect, the check on the intersection is performed on

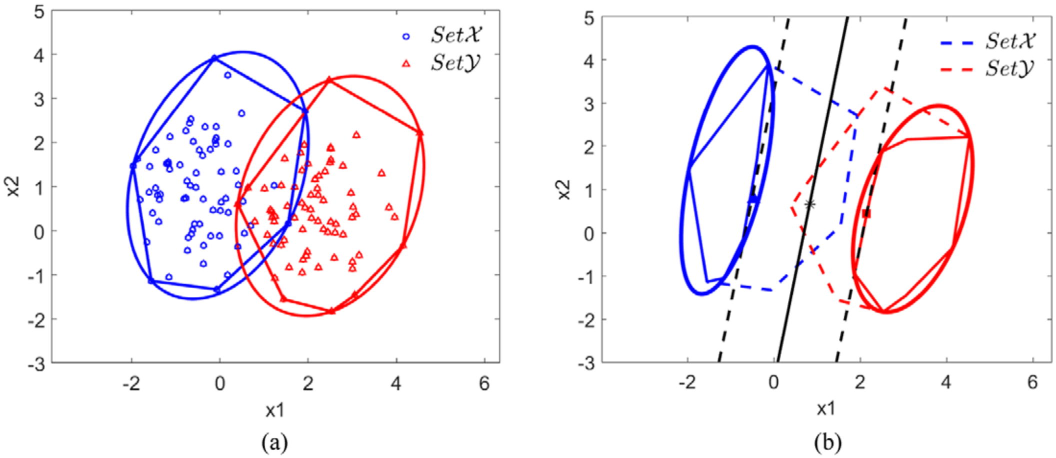


Figure 1. *(a) Intersecting CHs, and MVEs, of the two sets; (b) RCHs, and MVEs, that are nonintersecting (solid lines).*

https://doi.org/10.1017/dce.2024.41 Published online by Cambridge University Press

the MVEs that cover them. The RCHs intersect if their respective MVEs intersect. This approach is selected to minimize computational cost, especially for high-dimensional datasets. The MVEs shown in Figure 1b are the first partitions of the dataset. Having found these partitions, the points contained in them are removed from the dataset and SEP-C continues to find similar partitions on the remainder of the training data. If the dataset is linearly separable, then SEP-C terminates in the first iteration itself. For datasets with significant overlap, additional iterations are performed. SEP-C is guaranteed to terminate in a finite number of iterations, as either there are no more points to be partitioned or the I-D inequality is violated. The pseudo-codes of SEP-C and the RCH algorithm with time complexity analysis are described in detail in Niranjan and Rao (2023a).

2.3. Ellipsoidal partitioning algorithm

The pseudocode of ellipsoidal partitioning used in SEP-C is described in Algorithm 2.3; the notation $|\mathcal{X}|$ denotes the number of points in the set $\mathcal{X}$ and $\mathcal{X}\ \mathcal{Y}$ denotes the difference of sets $\mathcal{X},\mathcal{Y}$. This algorithm uses the RCH algorithm to find nonintersecting convex hulls of the two sets $\mathcal{X}$ and $\mathcal{Y}$ in each iteration, as explained in Section 2.2. The pseudocode for the RCH algorithm is presented in Niranjan and Rao (2023a) and omitted here in the interest of brevity. Algorithm 1 consists of two user-defined parameters: the integers $n > 0$, which denotes the dimension of the feature space, and $n_{\text{Imp}} \geq 0$, which denotes the number of permitted misclassifications.

Algorithm 1 Ellipsoidal partitioning.

```
$i \leftarrow 1$
$\mathcal{E}_X, \mathcal{E}_Y \leftarrow$ MVEs of $\mathcal{X}$ and $\mathcal{Y}$
$\mathcal{X}^+, \mathcal{X}^-, \mathcal{Y}^+, \mathcal{Y}^- \leftarrow \varnothing$
$\mathcal{E}^+, \mathcal{E}^- \leftarrow \varnothing$
$n_{\text{Imp}} \geq 0$                ▷ Number of points of one label allowed in the set of another
$n > 0$                ▷ Dimension of the feature space
while $|\mathcal{X}| > n$ and $|\mathcal{Y}| > n$ do                ▷ Ensure I-D condition is satisfied
    $\left(\mathcal{X}_i^+, \mathcal{Y}_i^-, \mathcal{E}_{X^+}, \mathcal{E}_{Y^-}\right) \leftarrow \text{RCH}(\mathcal{X}, \mathcal{Y})$                ▷ RCH Algorithm
    if $\mathcal{X}_i^+ = \varnothing\ \mathcal{Y}_i^- = \varnothing$ then
        Datasets $\mathcal{X}$ and $\mathcal{Y}$ cannot be separated further
        break
    else
        while $|\mathcal{Y} \in \mathcal{E}_{X^+}| \geq n_{\text{Imp}}$ do
            $\mathcal{X}_i^+ \leftarrow \text{RCH}\left(\mathcal{X}_i^+, \mathcal{Y}\right)$ ▷ Subset of $\mathcal{X}$ containing no greater than $n_{\text{Imp}}$ points of $\mathcal{Y}$; found by solving CP (2)
        end while
        while $|\mathcal{X} \in \mathcal{E}_{Y^-}| \geq n_{\text{Imp}}$ do
            $\mathcal{Y}_i^- \leftarrow \text{RCH}\left(\mathcal{Y}_i^-, \mathcal{X}\right)$ ▷ Subset of $\mathcal{Y}$ containing no greater than $n_{\text{Imp}}$ points of $\mathcal{X}$; found by solving CP (2)
        end while
        $\mathcal{X}^+ \leftarrow \left\{\mathcal{X}_i^+\right\}, \mathcal{Y}^- \leftarrow \left\{\mathcal{Y}_i^-\right\}$
        $\mathcal{E}_i^+ \leftarrow \text{MVE}\left(\mathcal{X}_i^+\right), \mathcal{E}_i^- \leftarrow \text{MVE}\left(\mathcal{Y}_i^-\right) > \text{CP (1)}$
        $\mathcal{E}^+ \leftarrow \left\{\mathcal{E}_i^+\right\}, \mathcal{E}^- \leftarrow \left\{\mathcal{E}_i^-\right\}$
        $\mathcal{X} \leftarrow \mathcal{X} \setminus \mathcal{X}_{\mathrm{i}}^+, \mathcal{Y} \leftarrow \mathcal{Y}\mathcal{Y}_{\mathrm{i}}^-$
        $\mathcal{E}_X, \mathcal{E}_Y \leftarrow$ MVEs of $\mathcal{X}$ and $\mathcal{Y}$
    end if
    $i \leftarrow i + 1$
end while
if $|\mathcal{X}|(\text{or}|\mathcal{Y}|) \geq n$ then
    $\mathcal{E}^+(\mathcal{E}^-) \leftarrow$ MVE of $\mathcal{X}(\mathcal{Y})$
```

```
29: else
30:     Consider points in $\mathcal{X}(\mathcal{Y})$ as individual ellipsoids
31: end if
```

The operation of SEP-C, based on Algorithm 1, is presented next, using the dataset irregularity of being skewed.

3. Handling class skew

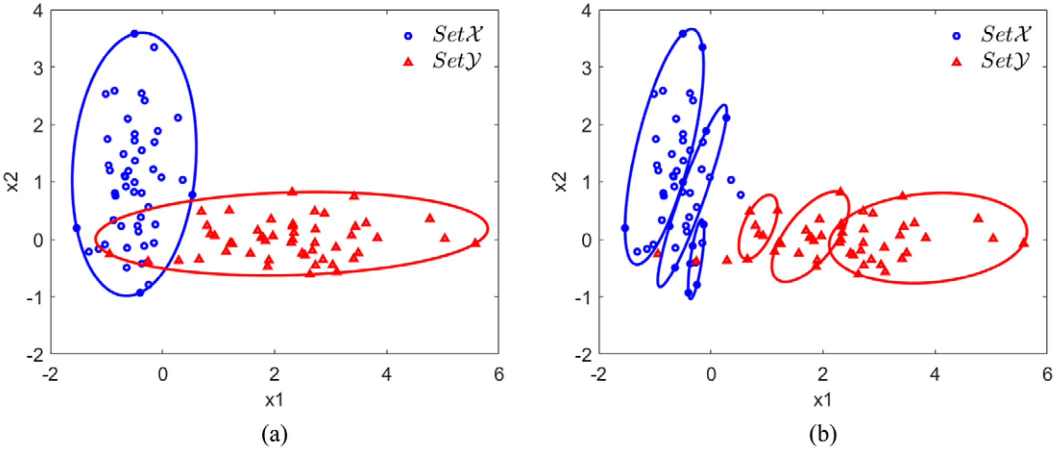
We first discuss how SEP-C is not influenced by class skew while acting as a supervised classifier. The ellipsoids found in each iteration of SEP-C are a basis for classification. Since, by construction, the ellipsoids contain mostly points of the same label (lines 14, 17, and 20 in Algorithm 1), if an unseen test point is now contained within one of them, it is assigned the label of the majority of the points in that ellipsoid; see Niranjan and Rao (2023a) for issues on calculating the trust score for such classification rules. In any iteration of SEP-C, the CP (2) is solved to find the points **c** and **d**. It can be observed that **c**, or **d**, is a convex combination of points in **X** *alone*, or **Y** *alone*, and not on any joint characteristics of the sets $\mathcal{X}$ and $\mathcal{Y}$. Further, if the solution to the CP (2) exists, then $\mathbf{c}, \mathbf{d}$, and the separating hyperplane, which is normal to the line joining these points, are also unique. The points **c** and **d** may lie in the region of intersection of the respective CHs, thus leading to points of both labels on either side of the obtained hyperplane. If the number of such points, which are in essence misclassifications, is less than the permitted number n_{Imp}, then SEP-C terminates in that iteration.

Since, for a "good" classifier, n_{Imp} should be low, SEP-C reduces the ellipsoids found in an iteration, using the RCH algorithm, but by iteratively changing the value of D. In this case, a reduced ellipsoid of one label is still independent of the ellipsoid found for the other label. This can be demonstrated by viewing the RCH algorithm akin to the classic Ellipsoid Algorithm discussed in Bland et al. (1981), where a smaller ellipsoid in a later iteration is found after performing a cut on the ellipsoid found in the earlier iteration. In our case, the cut is exactly the hyperplane found in that iteration. According to the Ellipsoid Algorithm, the center, x_k, and the properties of an ellipsoid, given by an SPD matrix $\mathbf{B}_k$, found in iteration k, expressed as $\mathcal{E}_k = \left\{x \in \mathfrak{R}^n \mid (x-x_k)^T \mathbf{B}_k^{-1} (x-x_k) \leq 1\right\}$, which is then cut by a line $a^T x = b$, leads to a smaller ellipsoid in iteration $(k+1)$ that are functions of a^T and x_k and B_k *alone*. In our case, let in iteration k, the ellipsoid $\mathcal{E}_k^+$, that contains a majority of points with label L_{+1}, also contain $n_k > n_{\text{Imp}}$ number of points with label L_{-1}. Now, in the next iteration, the smaller ellipsoid, $\mathcal{E}_{k+1}^+$, is such that it contains $n_{k+1} \leq n_{\text{Imp}}$ points with label L_{-1}; indeed, $\mathcal{E}_{k+1}^+$ may also now have fewer points of label L_{+1} (line 22 in Algorithm 1, where datapoints that are contained in an ellipsoid are removed). The key observation is that finding $\mathcal{E}_{k+1}^+$ is not dependent on the properties of the corresponding ellipsoid, $\mathcal{E}_k^-$, found for the other label. The partitions found by SEP-C for a synthetic dataset with points obtained from skewed distributions are shown in Figure 2 using $n_{\text{Imp}} = 2$. As can be seen, SEP-C is immune to class skew and the resulting ellipsoids contain no more than two points of the other label.

4. Identifying small disjuncts

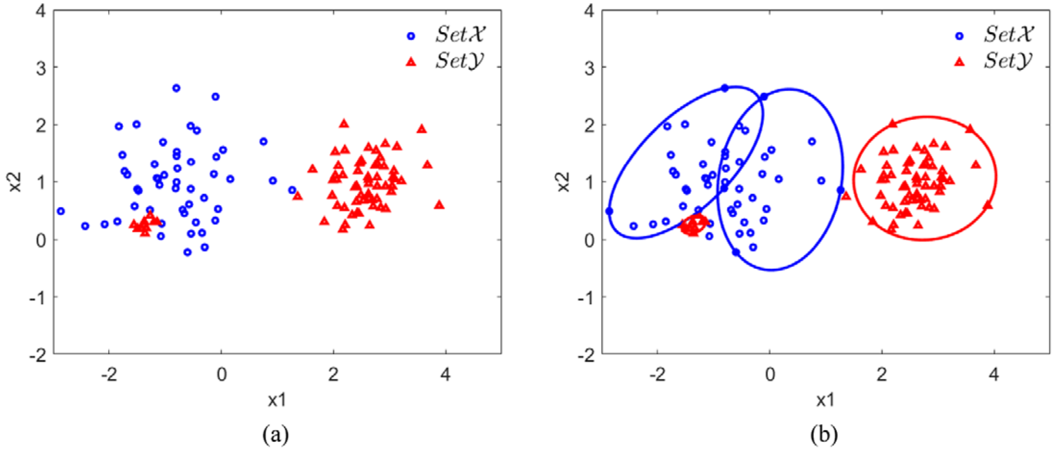
The use of the SEP-C method to detect small disjuncts or rare cases is now discussed. As finding such rare cases is an exploratory exercise, the entire dataset is used for partitioning. Ideally, SEP-C should isolate these rare cases in their own ellipsoids. Consider the 2D synthetic dataset shown in Figure 3a, where a few points (10 of them) of set $\mathcal{Y}$ have characteristics that are different from the majority of points belonging to that set; it is evident that a single hyperplane will not be able to isolate this small cluster, thus leading to errors in classification. Indeed, kernels, such as radial basis functions may be used, but these require considerable tuning of the hyperparameters. As can be seen in Figure 3b, SEP-C is able to isolate these points in their own ellipsoid. For this case, 2 iterations, with $n_{\text{Imp}} = 5$, are sufficient leading to 2 ellipsoids for each label.

https://doi.org/10.1017/dce.2024.41 Published online by Cambridge University Press



(a) (b)

Figure 2. *Partitioning a skewed 2D dataset.*



(a) (b)

Figure 3. *Partitioning 2D dataset with a small disjunct.*

It can happen that these rare cases are in proximity of points of another label or if the number of such rare cases is lesser than the dimension of the dataset; in the latter case, an MVE cannot be found exclusively as the I-D inequality fails. In both cases, choosing an appropriate value for n_{Imp} aids in identifying these rare, but non-trivial number of points. Suppose a small disjunct is defined as comprising of $5 < n_{\text{SD}} < 15$ number of points. Now, by choosing $n_{\text{Imp}} \geq n_{\text{SD}}$, SEP-C can find ellipsoids of one label to also contain n_{Imp} points, or fewer, of the other. Suppose an ellipsoid does contain $n_{\text{SD}} \geq n$ points, where n is the dimension of the feature space. Now, an MVE can be wrapped around these n_{SD} points and checked if it is contained completely in the larger ellipsoidal partition. If so, these points of the other class suggest a subconcept, which can be a potential small disjunct. The number n_{SD} now gives the *coverage* of the disjunct. In higher dimensional datasets, where $n > n_{\text{SD}} > 15$, though an MVE cannot be found for the disjunct, density-based measures can be used to establish if the n_{SD} points qualify as a *small disjunct* or are just noise.

4.1. Results of identifying small disjuncts

SEP-C is applied on the Vote, Pima Indian Diabetes (PID), and Chess Endgame datasets (Dua and Graff, 2017), to identify small disjuncts. The 3 partitions obtained for the Vote dataset, with $n_{\text{Imp}} = 17$, did *not* contain any small disjuncts in any of the partitions. On the other hand, SEP-C partitioned the PID dataset

https://doi.org/10.1017/dce.2024.41 Published online by Cambridge University Press

Table 1. *Ellipsoidal Partitions of the PID dataset, and their coverage, for different values of*n_{Imp}

	Number of ellipsoids	1	2	3	4	5	6	7	8	9	10	11	12	13
$n_{\mathrm{Imp}}=5$	Class 0	58 (2)	73 (3)	64 (4)	27 (3)	44 (2)	25 (3)	15 (3)	23 (3)	35 (4)	24 (4)	21 (4)	27 (4)	22 (3)
	Class 1	58 (2)	73 (3)	64 (4)	27 (3)	44 (2)	25 (3)	15 (3)	23 (3)	35 (4)	24 (4)	21 (4)	27 (4)	22 (3)
$n_{\mathrm{Imp}}=10$	Class 0	**64 (9)**	**150 (9)**	36 (5)	32 (6)	34 (6)	**26 (9)**	35 (7)	**52 (9)**					
	Class 1	57 (4)	37 (5)	35 (7)	14 (6)	18 (7)	**40 (8)**	22 (5)	**19 (8)**					

into 8 ellipsoids for both labels (`no diabetes` and `diabetes`), with n_{Imp}=10, and 6 of these ellipsoids contained possible small disjuncts of sizes 9 and 8, respectively; the MVEs of these small disjuncts did not intersect with other ellipsoids of the same label, thus indicating that they are subconcepts of points with that label. The ellipsoidal partitions obtained for $n_{\mathrm{Imp}}=5,10$ are listed in Table 1. The number in the brackets indicates the number of points of the other class in each ellipsoid. The Chess Endgame dataset, with 36 features and $n_{\mathrm{Imp}}=37$, was partitioned into 5 ellipsoids each for both labels. Three ellipsoids have potential small disjuncts of sizes 5, 11, and 5.

5. Identifying dataset shift

The ellipsoids obtained by partitioning a dataset using SEP-C, such as shown in Figures 2b and 3b, have different orientations and centers. Each of these ellipsoids, obtained by solving the CP (1), can be expressed in the form $\mathcal{E}=\left\{x\in\Re^n \mid (x-\mu)^T\Sigma^{-1}(x-\mu)\leq 1\right\}$, where μ is the centre of the ellipsoid and the SPD matrix Σ admits the eigenvalue decomposition $\Sigma=\mathbf{Q}\Lambda\mathbf{Q}^T$, where $\mathbf{Q}$ are the orthogonal eigenvectors and the Λ is the diagonal matrix of eigenvalues. By expressing the part of the dataset as being contained in the ellipsoid, the probability density function of those points can be computed using a continuous multivariate Gaussian distribution, where Σ is the covariance matrix and μ is the mean. In addition, by a change of variables based on the eigendecomposition of Σ, these points can be transformed to being generated from n i.i.d. univariate Gaussians. In Figure 4a, ellipses obtained by applying SEP-C on a 2D synthetic dataset and the transformation of one of them (red in Figure 4a) to be centered at the origin and oriented along the axes (magenta in Figure 5) can be seen.

The role of SEP-C in detecting a shift in the dataset now becomes clear. Suppose the dataset undergoes a shift in an unknown way and new samples emerge owing to this shift, SEP-C expands the ellipsoids

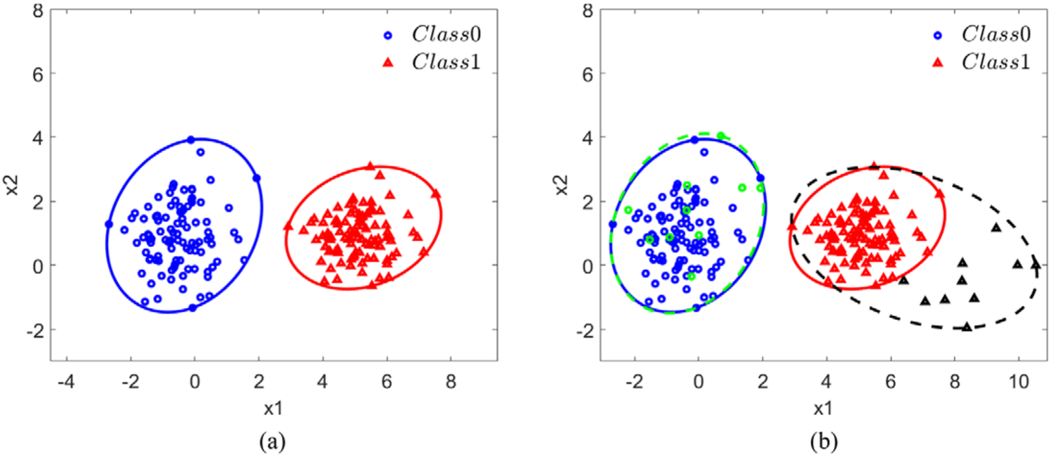


Figure 4. *SEP-C results for 2D dataset with OOD test samples.*

https://doi.org/10.1017/dce.2024.41 Published online by Cambridge University Press

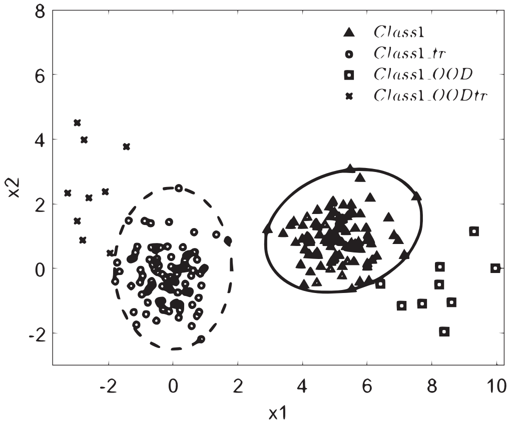


Figure 5. *Class 1 points and MVE transformed to origin with OOD points.*

closest to these points to cover them and assign a label; note that depending on the nature of the shift, it could be any of the ellipsoids in Figures 2b and 3b that undergos this expansion. It should be mentioned that to detect a shift, the act of classification takes a backseat in comparison with identifying the occurrence of this shift. Now, as i.i.d. univariate Gaussians can be computed using the ellipsoids, standard tests to detect shift in distribution can be applied, for example, the KS test, Dodge (2008). As is known, this test is applicable to continuous distributions and indicates a change by accepting or rejecting a hypothesis at some level, say 95%.

For example, let $F_0 = \mathcal{N}\left(0, \lambda_{i0}^{-0.5}\right)$ be the zero-mean univariate Gaussian along the direction i with standard deviation $\lambda_{i0}^{-0.5}$ obtained from applying SEP-C on the original data. Now, if there indeed is a shift, either mean or covariate, it is clear that the new data points will most likely not belong to F_0; the KS statistic computed using some confidence interval (https://in.mathworks.com/help/stats/kstest.html) will indicate if the null hypothesis (data comes from F_0) should be accepted or rejected.

5.1. Results of identifying dataset shift—2D synthetic dataset

Consider the synthetic, 2D, linearly separable dataset with points of two classes, 0 and 1, as shown in Figure 4a. The overlap check built into SEP-C indicates that the MVEs that cover these points do not overlap. To detect a shift in data, 10 in-distribution test samples for Class 0 (green) and 10 OOD test samples for Class 1 (black) are synthetically generated and shown in Figure 4b. As all Class 0 points are in-distribution, there is no (significant) change in the corresponding MVE. On the other hand, the MVE for Class 1 undergoes considerable expansion, marked in red and black in Figure 4b, respectively. Running the KS test using the properties of the distribution computed using the original ellipsoid indicates that the null hypothesis should be rejected, that is, these samples do not belong to the same distribution as the training data, in turn, indicating a shift in data. When the KS test fails for a sufficient number of new samples, the user may choose to disregard the old data and perform partitioning on the new.

It should be remarked that the type of test applied to detect change in distribution is not crucial here; the OOD samples could also have been detected by computing the Mahalanobis distances or applying variants of the KS test. The key aspect is that SEP-C provides a way by which such a change can be measured, independent of the distribution of the new data, or the old, or the class of points that undergo this shift.

5.2. Results of identifying dataset shift—MNIST dataset

SEP-C is applied to detect OOD samples using the MNIST dataset, Deng (2012). As an illustration, first, 1000 images from the MNIST dataset containing 97 images of digit 0 and 903 of digits 1–9 are chosen. Since the images are of size 28×28, in order to reduce the computation time, the image size is reduced to

https://doi.org/10.1017/dce.2024.41 Published online by Cambridge University Press

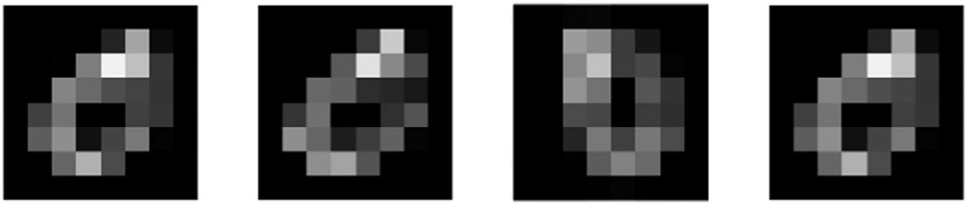


Figure 6. *Digit 0 image of reduced size with speckle, rotation, and filter distortions.*

8×8. By expressing each image as a vector of size 64×1 and adopting the one-versus-all approach, the SEP-C algorithm yielded an ellipsoid containing 92 images of digit 0 and a nonintersecting ellipsoid containing 84 remaining digits; denote the ellipsoid containing digit 0 images as $\mathcal{E}_0$. The algorithm terminated after the first iteration since 5 of digit 0 images were left and hence, as the I-D inequality fails, further ellipsoids cannot be found; also not that 819 images of digits 1–9 also remain. However, as the focus of this section is to show how the ellipsoidal parameters change with dataset shift, it is sufficient to analyze the ellipsoid that contains images of digit 0 from the training dataset.

To create a dataset shift, the next 500 images (also used as a testing set) are subjected to the following distortions: i. addition of speckle noise; ii. rotation by 60°; and iii. distortion using the *disk* filter implemented using the `imfilter` function in Matlab; such distortions are also employed in Rabanser et al. (2019). These distortions are shown in Figure 6 for the digit 0.

The testing set contains 43 images of digit 0. As mentioned in the introduction to SEP-C, distances of a test point from all ellipsoids are computed the ellipsoid closest to this test point is expanded to cover it. For the distorted digit 0 images in the test set, for each type of distortion, several of these images were the closest to $\mathcal{E}_0$ and hence, this ellipsoid had to be expanded to cover them; denote this ellipsoid as $\mathcal{E}'_0$. It is observed that the Euclidean distances of the centers of $\mathcal{E}_0$ and $\mathcal{E}'_0$, for the three types of distortion, are 0.14, 0.19, and 0.11, respectively. Note that since the elements of the 64×1 vector, that define each image, are contained between 0 and 1 (as the images are grayscale), such changes in the locations of the centers are significant. The ratios of the eigenvalues of the matrices defining $\mathcal{E}'_0$ and $\mathcal{E}_0$ also reflect a change with the different distortions; define this ratio as $r_i = \frac{\lambda_i(\mathcal{E}'_0)}{\lambda_i(\mathcal{E}_0)}, \quad i = 1, \cdots, 64$. It is observed that the ratios lie in the range $r_i \in [0.1, 1.02]$ for the speckle distortion; $r_i \in [0.03, 1.03]$ for the rotation distortion; and $r_i \in [0.38, 1.02]$ for the filter distortion. As can be seen, there is indeed a reduction in the magnitudes of the eigenvalues, indicating an increase in the corresponding semi-axes lengths and in turn, that the ellipsoid $\mathcal{E}'_0$ is larger and different than $\mathcal{E}_0$; thus, SEP-C is able to detect a shift in the data.

6. Conclusion

This paper presented the application of the SEP-C algorithm for identification of small disjuncts and shift in a dataset. As has been shown, this approach does not require additional preprocessing techniques when underlying data distribution is unknown, skewed, shifted or imbalanced. In all the cases, the proposed method uses a single user-defined hyperparameter, which controls the number of misclassifications to derive the ellipsoidal partitions. SEP-C can fail to capture small disjuncts when the number of samples is less than the dimensionality of the dataset since I-D inequality constraint restricts the formation of an MVE. Applying density-based measures in such cases for determination of small disjuncts could be a direction in which future study can be conducted. Robust statistical tests can be conducted to detect change in distributions as well.

Data availability statement. This work has used 2D synthetic data and publicly available datasets in the UCI repository and the MNIST datasets that have been cited.

Acknowledgments. Ranjani Niranjan would like to thank Prateeksha Foundation for providing financial support for her doctoral program at IIIT-Bangalore.

Author contribution. Conceptualization: Sachit Rao; Ranjani Niranjan. Methodology: Sachit Rao; Ranjani Niranjan. Data curation: Ranjani Niranjan. Data visualisation: Ranjani Niranjan. Writing original draft: Ranjani Niranjan; Sachit Rao. All authors approved the final submitted draft.

Funding statement. This research was not supported by any research funding.

Competing interest. None.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Barrows J, **Radu V**, **Hill M and Ciravegna F** (2021) Active learning with data distribution shift detection for updating localization systems. Paper presented at 2021 International Conference on Indoor Positioning and Indoor Navigation (IPIN), 29 November-02 December, Lloret de Mar, Spain.

Bennett KP and Bredensteiner EJ (2000) Duality and geometry in SVM classifiers. In Langley P (eds), *Proceedings of the Seventeenth International Conference on Machine Learning, ICML'00.* San Francisco, CA: Morgan Kaufmann Publishers, 57–64.

Bland RG, **Goldfarb D and Todd MJ** (1981) The ellipsoid method: A survey. *Operations Research*, *29*(6):1039–1091.

Boyd S and Vandenberghe L (2004) *Convex optimization*. Cambridge, UK: University Printing House, Cambridge University Press.

Boyd S and Vandenberghe L (2018) *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*. Cambridge, UK: University Printing House, Cambridge University Press.

Das S, **Datta S and Chaudhuri BB** (2018) Handling data irregularities in classification: Foundations, trends, and future challenges. *Pattern Recognition*, *81*:674–693.

Deng L (2012) The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, *29* (6):141–142.

Dodge Y (2008) *Kolmogorov–Smirnov Test*. Springer New York, 283–287.

Kelly, M, **Longjohn, R**, **Nottingham, K** (2023) The UCI Machine Learning Repository, https://archive.ics.uci.edu

Guerin J, **Delmas K**, **Ferreira R and Guiochet J** (2023) Out-of-distribution detection is not all you need. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(12):14829–14837.

Holte RC, **Acker LE and Porter BW** (1989) Concept learning and the problem of small disjuncts. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence, IJCAI'89*, vol *1*. San Francisco, CA. Morgan Kaufmann Publishers, 813–818.

Jo T and Japkowicz N (2004) Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter*, *6*(1):40–49.

Kong Q and Zhu Q (2007) Incremental procedures for partitioning highly intermixed multi-class datasets into hyper-spherical and hyper-ellipsoidal clusters. *Data & Knowledge Engineering*, *63*(2):457–477.

Lee K, **Lee K**, **Lee H and Shin J** (2018) A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N and Garnett R (eds), *Advances in Neural Information Processing Systems*. Curran Associates. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada.

Li A, **Boyd A**, **Smyth P and Mandt S** (2021) Detecting and adapting to irregular distribution shifts in bayesian online learning. In Ranzato M, Beygelzimer A, Dauphin Y, Liang P and Vaughan JW (eds), *Advances in Neural Information Processing Systems*, vol *34*. Curran Associates, 6816–6828. 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Niranjan R and Rao S (2023a) Classification with trust: A supervised approach based on sequential ellipsoidal partitioning. *IEEE Transactions on Knowledge and Data Engineering,* 1–14

Niranjan R and Rao S (2023b) Handling small disjuncts and class skew using sequential ellipsoidal partitioning. In Maji P, Huang T, Pal NR, Chaudhury S and De RK (eds), *Pattern Recognition and Machine Intelligence: 10th International Conference, PReMI 2023, Proceedings*. Springer-Verlag, 80–88.

Prati RC, **Batista GEAPA and Monard MC** (2004) Learning with class skews and small disjuncts. In Bazzan, ALC and Labidi, S, editors, *Advances in Artificial Intelligence – SBIA 2004*, pages 296–306. Berlin/Heidelberg: Springer.

Rabanser S, **Günnemann S and Lipton Z** (2019) Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, *32*:1396–1408.

Sastry CS and Oore S (2020) Detecting out-of-distribution examples with Gram matrices. In Hal D and Singh A (eds), *Proceedings of the 37th International Conference on Machine Learning*, Cambridge, MAL PMLR, 8491–8501. JMLR.org.

Sun P and Freund RM (2004) Computation of minimum-volume covering ellipsoids. *Operations Research*, *52*(5):690–706.

Tukey JW (1977) Exploratory Data Analysis. New York, NY: Addison-Wesley.

Weiss GM (2010) *The Impact of Small Disjuncts on Classifier Learning*. Boston, MA: Springer US, 193–226.

Weiss GM and Hirsh H (2000) A quantitative study of small disjuncts. *AAAI/IAAI*, *2000*(665–670):15.

Cite this article: Niranjan R and Rao S (2024). Handling shift and irregularities in data through sequential ellipsoidal partitioning. *Data-Centric Engineering*, 5, e37. doi:10.1017/dce.2024.41