The Value of Diversity for Dealing with Concept Drift in Class-Imbalanced Data Streams

Chun Wai Chiu
Baxall Construction Ltd.
Paddock Wood, United Kingdom
mchiu@baxallconstruction.co.uk

Leandro L. Minku
School of Computer Science, *University of Birmingham*
Birmingham, United Kingdom
L.L.Minku@bham.ac.uk

***Abstract*—Concept drift and class imbalance are critical challenges in real-time data stream learning. Existing ensemble methods use homogeneous diversity (models for the same concept) to tackle these challenges but often overlook heterogeneous diversity (models from different concepts), which could improve adaptation, especially with scarce minority data. This paper provides the first analysis of when and why each type of diversity is beneficial for class-imbalanced data streams. To enable this analysis, we introduce CDCMS.CIL, a novel class imbalance learning framework for leveraging heterogeneous diversity. Experiments based on 80 artificial and 9 real-world data streams show that heterogeneous diversity can significantly aid concept drift handling in highly imbalanced scenarios, while homogeneous diversity is better during stable periods. These findings provide crucial guidance for designing robust ensembles for drifting class imbalanced data streams.**

***Index Terms*—data stream learning, class imbalance, ensemble diversity.**

I. Introduction

The increasing volume of high-speed data streams necessitates adaptive data stream learning algorithms. Real-world data streams present two major challenges for machine learning systems. First, concept drift occurs when the underlying data distribution changes over time, making existing models obsolete [1]. Second, class imbalance arises when one class significantly outnumbers others, leading to poor recognition of minority class examples [2]. Examples include fraud detection (rare fraudulent transactions, evolving fraud patterns) [3], [4], intrusion detection (scarce malicious traffic, evolving attack methods) [5], and medical diagnosis from sensor data (infrequent critical health events, changing patient conditions) [6]. Addressing both challenges simultaneously is particularly difficult due to the scarcity of minority class data for updating models after drifts [7].

To address these challenges, existing ensemble approaches typically leverage *homogeneous diversity*, creating diverse models that all represent the *same concept* with techniques like Bagging or Boosting [8]–[10]. However, recent studies suggest that *heterogeneous diversity* (maintaining models of *different concepts*) may better address sudden and recurring drifts [11]. In drifting class-imbalanced data streams, this could aid post-drift performance recovery by leveraging past knowledge when new minority examples are scarce. Yet, this scarcity could also hinder the identification of relevant past models. Without thorough research, the value of heterogeneous diversity for class-imbalanced data stream learning is still uncertain.

This study provides the first investigation into whether, when and why each type of diversity (homo/heterogeneous) is beneficial for class-imbalanced data streams. Given the lack of heterogeneous diversity approaches in this context, we propose a novel framework designed for class-imbalanced data streams to support the analysis. This framework facilitates strategies for maintaining a heterogeneously diverse memory of models in class-imbalanced data streams, including when to store new models, replace existing models, and recover past models for predictions. It leverages relevant past knowledge to adapt to multiple types of drift in class-imbalanced scenarios while overcoming limitations of existing diversity-based approaches such as excessive memory usage [9], [12] and reliance solely on a recent window of data [11], [13]. This study answers the following research questions:

- **RQ1:** How to best leverage heterogeneous diversity for class-imbalance data stream learning?
- **RQ2:** How helpful is heterogeneous diversity in the context of imbalanced data streams (RQ2a)? When and why is it beneficial or prone to failure (RQ2b)?
- **RQ3:** How helpful is homogeneous diversity in the context of imbalanced data streams (RQ3a). When and why is it beneficial or prone to failure (RQ3b)?

RQ1 is addressed through our proposed framework and an investigation of its variations (e.g., weighting metrics, resampling). Our findings suggest that time-decay G-Mean and oversampling are often most effective for the proposed framework. RQ2 and RQ3 are addressed by an extensive evaluation of 9 approaches, including 7 existing diversity-based approaches, 2 baselines, and our novel framework, on 80 artificial and 9 real-world data streams. Results indicate that heterogeneous diversity is particularly beneficial for handling concept drift in recurring and severely imbalanced scenarios, while homogeneous diversity excels when concepts are stable.

Overall, the main contributions of this paper are:

- The first systematic analysis of when and why different types of diversity (homogeneous/heterogeneous) are helpful in the context of drifting class-imbalanced data streams. Among other findings, heterogeneous diversity

This work was partly funded by a School Scholarship provided in support of the EPSRC grant EP/R006660/2.

improves concept drift adaptation especially for recurring and severe drifts by effectively use of past relevant knowledge. Homogeneous diversity is most effective during stable concepts or mild drifts.
- The introduction of a novel heterogeneous diversity framework (CDCMS.CIL) designed to address the gap in drifting class-imbalanced data streams, which lacks such kind of approaches. The approach obtained competitive predictive performance against representative diversity-based approaches and baselines on a wide range of artificial and real-world datasets.

The rest of this paper is organised as follows. Section II presents the problem formulation. Section III discusses related work on diversity and approaches for drifting class-imbalanced data streams. Section IV presents the proposed framework. Section V presents the experimental study and results. Section VI concludes the study and proposes future directions.

II. Problem Formulation

In supervised learning, a data stream $DS = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots\}$ is a potentially infinite sequence of examples arriving in chronological order, where $\mathbf{x}_t$ is a feature vector and y_t is the class label [1], [14]–[16]. The underlying joint probability distribution $D_t = P(\mathbf{x}_t, y_t)$ of DS is referred to as the *concept* [17]. Typically, the concept of a stream changes over time $(\exists t, D_t \neq D_{t+1})$, which is referred to as *concept drift* [1], [15], [16], and can render existing models obsolete. For each new example $(\mathbf{x}_t, y_t)$, we aim to update an ensemble learning system to generalise to D_t (*online data stream learning*), potentially with limited access to past examples. We assume the data stream is *class-imbalanced*, i.e., the probability of seeing examples from one class is much lower than from the others $(\exists i, j, P_t(y^i) << P_t(y^j))$ [7], [18]. This can significantly impair the ability to recognise minority class examples.

III. Related Work

A. Diversity for Drifting Class-balanced Data Streams

Existing work has shown that high levels of homogeneous diversity help concept drift adaptation [19], lower levels are better for stable periods [19], and heterogeneous memories of past models representing different concepts can support efficient learning of recurring concepts [13]. Several ensemble data stream learning approaches have been proposed to exploit different kinds of diversity for dealing with concept drift, including approaches likely to result in homogeneous diversity (e.g., Diversity for Dealing with Drift (DDD) [20], Adaptive Diversified Ensemble Selection Classifier (ADES) [21] and Accuracy-and-Diversity-based Ensemble (ADE) [22]), and heterogeneous diversity (e.g., Diversity Pool (DP) [13], Diversity and Transfer-based Ensemble Learning (DTEL) [23] and Concept Drift handling based on Clustering in the Model Space (CDCMS) [11]). Besides, some drift detection methods, e.g., Diversity Measure Drift Detection Method (DMDDM) [24], have also been proposed based on diversity. However, none of the aforementioned studies were specifically designed for class-imbalanced data streams, leaving the role of diversity in this context undefined. It remains unclear if diversity would play a similar role in class-balanced and imbalanced streams, primarily due to the challenges posed by the scarcity minority class examples. This scarcity may complicate recovery from drifts in homogeneous ensembles and the identification of distinct concepts in heterogeneous ensembles.

B. Approaches for Drifting Class-Imbalanced Data Streams

Addressing the joint challenge of concept drift and class-imbalance is gaining increasing attention [25]. Most existing ensemble methods rely on *homogeneous diversity*, where ensemble members are diversified but aim to represent the current concept. Oversampling and Undersampling Online Bagging (OOB, UOB) [8] are pioneering homogeneous ensemble approaches based on resampling whose performance remains competitive against recent approaches [2]. Their variations $\mathrm{OOB_d}$ and $\mathrm{UOB_d}$ with drift detection methods can be used to deal with real drifts [18]. Dynamic Weighted Selective Ensemble (DWSE) [26] uses a dual-window system to train and dynamically weight base learners on recent and "hard" minority instances. Underperforming base learners are replaced by new ones to adapt to drift, rendering DWSE a homogeneous diversity approach. Noisy-Sample-Removed Undersampling (NUS) [27] clusters and removes less informative majority instances before training. Its application in generating varied datasets for ensembles of similar classifiers makes it a homogeneous diversity approach.

Other methods involve generating synthetic minority class examples. Continuous-SMOTE (C-SMOTE) [28] adapts the offline SMOTE technique [29] for data streams using an adaptive window but can cause memory issues in the absence of concept drift. Very Fast Continuous-SMOTE (VFC-SMOTE) [30] addressed this by using a "sketch" data structure to summarise past example and generating synthetic minority examples SMOTE with Online Bagging (SMOTE-OB) [10] integrates this strategy into OnlineUnderOverBagging [9], benefiting from three data-level re-balancing strategies. Nevertheless, all these approaches may generate noisy examples. Alternatively, cost-sensitive learning is used by approaches like Cost-sensitive Adaptive Random Forest (CSARF) [31], an online ensemble that assigns weights to base learners based on the Matthews Correlation Coefficient to address class imbalance. Its ARF foundation inherently uses homogeneous diversity.

Despite these developments, all approaches mentioned rely on homogeneous diversity. The specific benefits of using diversity for drift adaptation in class-imbalanced streams remain largely unexamined, and none of these methods have explored the use of heterogeneous diversity.

IV. Proposed Framework

This section proposes the first heterogeneous diversity framework for drifting class-imbalanced data streams – Concept Drift Handling Based on Clustering in the Model Space for Class-Imbalanced Learning (CDCMS.CIL) [1]. Section IV-A presents an overview of CDCMS [11], as a prerequisite to CDCMS.CIL. Section IV-B motivates and gives an overview of the proposed CDCMS.CIL. Sections IV-C to IV-E explain its diversity-based memory management strategies specifically designed for class-imbalanced data stream learning.

A. Concept Drift Handling Based on Clustering in the Model Space (CDCMS)

CDCMS [11] is an ensemble approach that employs diversity to address multiple types of concept drift in class-balanced data streams. It comprises four key elements: a main ensemble, a memory (repository) of past base learners, a drift detector, and a recent window of examples. The *main ensemble*, composed of online learning base learners, predicts new instances through a weighted majority vote. The *memory* stores a diverse set of past base learners which are not active in prediction but can be useful if drift occurs. To deal with gradual drifts, CDCMS periodically introduces a new base learner to replace the least-performing one in the main ensemble. To deal with abrupt drifts, it employs a drift detector and a clustering in the model space strategy.

Upon drift detection, CDCMS clusters all past base learners in the memory (a.k.a., clustering in the model space) based on their predictions on the recent *window of examples*. Each cluster represents a potentially different concept. It then forms a heterogeneously diverse auxiliary ensemble (NH) with a representative from each cluster. NH is then make predictions alongside the main ensemble (NL) to mitigate post-drift performance drops by leveraging various past concepts. To form a new main ensemble post-drift, CDCMS first creates a new base learner c_n to learn alongside the old main ensemble upon drift detection. At the b-th time steps since the drift detection, c_n is clustered with past learners to find those relevant to the current concept, which are then reinstated to form the new main ensemble. If c_n forms a unique cluster, the post-drift concept is deemed novel, and the new main ensemble is built from scratch, initially with only c_n.

To optimise the number of concepts within a limited memory, CDCMS employs a diversity-based memory management strategy. When a base learner c must be moved from the main ensemble to a full memory, CDCMS uses Yule's Q-statistics to identify the past base learner c' that is most similar to c, based on a similarity threshold θ. The learner (c or c') trained on more examples is retained. If no sufficiently similar past model exists, c is discarded.

B. Overview and Motivation of the Proposed Approach

CDCMS uses a recent window of examples for model retrieval via model space clustering upon concept drift detection. However, this window often lacks sufficient examples to represent the decision regions of the current or past concepts. This issue is exacerbated when minority class examples are scarce, hindering the similarity analysis between base learners during clustering in the model space. Fig. 1 illustrates this issue. Intuitively, based on decision areas, c_1 and c_2 should be clustered together, as should c_3 and c_4. Nonetheless, clustering based on the predictions to the recent window incorrectly groups c_1, c_2, and c_3 together since they all classify the window's examples as the majority class. Base learner c_4 ends up isolated, as it classifies some examples as the minority class. This strategy overlooks the fair consideration of each base learner's decision areas across both classes due to limited coverage of examples, a problem that is exacerbated in class imbalanced problems.

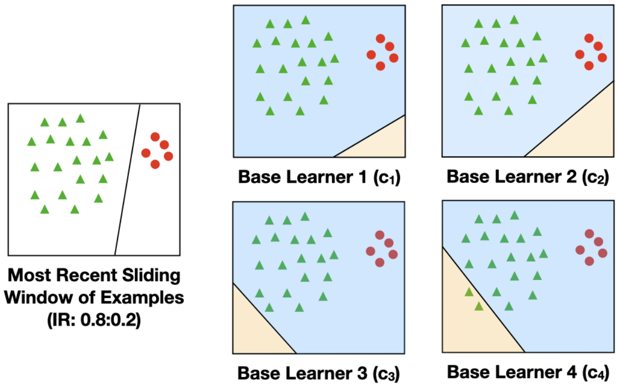


Fig. 1. Comparing similarities between base learners based on the most recent sliding window of examples. Green triangles represent the majority class examples and red circles the minority class examples in the sliding window. Blue and orange areas represent the learnt decision areas of the majority and minority class, respectively, for base learners c_1, c_2, c_3, and c_4.

To address this limitation, CDCMS.CIL builds a sparse representation of training examples for each base learner. This representation captures the spatial distribution of examples from both classes as learning progresses (Step 1 of Fig. 2). This strategy considers each base learner and class individually, more accurately capturing decision areas and preventing minority class examples to be confused with noise. This sparse representation, encompassing all examples seen so far, better represents various feature space regions than a recent window. Upon drift detection, base learners make predictions to synthetic examples created from this representation to determine which base learner to recover from memory (Steps 2-3 of Fig. 2). This set of synthetic examples is called the *projection set*. Clustering in the model space is then performed based on these predictions (Step 4 of Fig. 2).

C. Sparse Representation of Past Data Stream Examples

To create the sparse representation, each base learner in CDCMS.CIL is coupled with two stream clustering methods ($SC[]$) that operate in the feature space, summarising training examples from different classes as *micro-clusters*. Stream clustering methods capable of adapting to concept drift and constraining the number of micro-clusters are applicable. This study employs CluStream [32], as it provides the necessary capabilities. A micro-cluster (mc) typically includes the count of data points and vectors representing the linear and squared

[1] The source code of CDCMS.CIL is available at: https://github.com/michaelchiucw/CDCMS.CIL/tree/main/Implementation

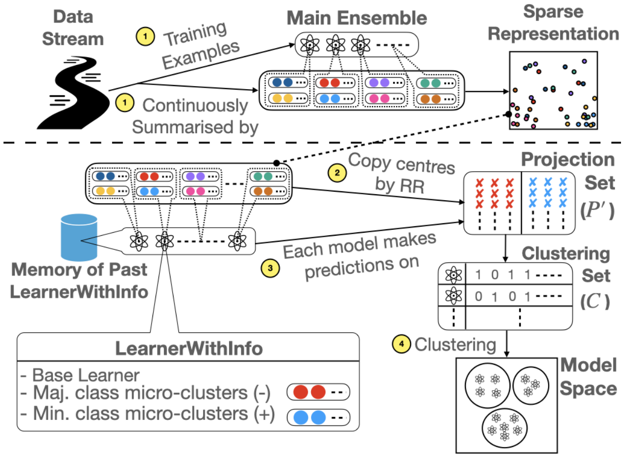


Fig. 2. Illustration of Clustering in the Model Space Strategy for Class-Imbalanced Learning

sums of locally close examples in the feature space [33], [34]. As defined in [32], $mc = (\overline{CF2^x}, \overline{CF1^x}, CF2^t, CF1^t, n)$. It summarises n d-dimensional points $\{\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_n}\}$, where

$$\overline{CF2^x} = \left(\sum_{i=1}^{n} (x_{t_i,1})^2, \ldots, \sum_{i=1}^{n} (x_{t_i,d})^2 \right),$$

$$\overline{CF1^x} = \left(\sum_{i=1}^{n} x_{t_i,1}, \ldots, \sum_{i=1}^{n} x_{t_i,d} \right),$$

$$CF2^t = \sum_{i=1}^{n} (t_i)^2, \quad CF1^t = \sum_{i=1}^{n} t_i.$$

and t_i denotes the timestamp of the i-th example within mc. Hence, each base learner c_i in CDCMS.CIL is associated with

$$MC_{c_i}^{class} = \{mc_1, \ldots, mc_{K_i}\},$$

where K_i is the number of micro-clusters associated to c_i and $class \in \{0, 1\}$. The overall sparse representation is

$$S = \{MC_{c_1}^0 \cup MC_{c_1}^1 \cup \ldots \cup MC_{c_L}^0 \cup MC_{c_L}^1\},$$

where $c_1, \ldots, c_L$ are all base learners in the system.

D. Clustering in the Model Space for Class-Imbalanced Data Stream Learning

The centres of the micro-clusters composing the sparse representation of the data stream can be seen as synthetic examples forming a set that we call the "projection set" P:

$$P = \{\Upsilon_{c_1}^0 \cup \Upsilon_{c_1}^1 \cup \ldots \cup \Upsilon_{c_L}^0 \cup \Upsilon_{c_L}^1\},$$

where $\Upsilon = \{\mu_i | \forall \mu_i \in MC\}$ is a set of micro-clusters' centres, and $\mu = \frac{\overline{CF1^x}}{n}$ is the centre of a given micro-cluster.

The clustering in the model space strategy proposed in this study clusters the base learners' predictions on P. As P covers the decision areas of both classes learnt by each base learner, this strategy ensures a fair consideration of them. Nonetheless, disparities in the number of micro-clusters of each base learner and class may introduce bias. For instance, some base learners may have more micro-clusters in the majority class and / or have accrued more due to prolonged exposure to the data stream. To address this, CDCMS.CIL employs a round-robin method to oversample micro-cluster centres within each set ($MC_{c_i}^{class}$), so that the total synthetic examples for each $MC_{c_i}^{class}$ equals $K_{max} = max(K_i)$. We refer to this resampled projection set as P', i.e.,

$$P' = \{RR(\Upsilon, K_{max}) | \forall \Upsilon\} \in P\}.$$

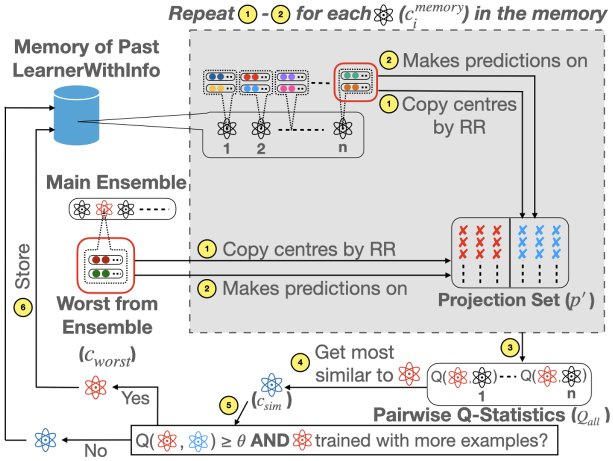

Repeat ①-② for each (c_i^memory) in the memory
Memory of Past LearnerWithInfo
② Makes predictions on
① Copy centres by RR
1
2
n
Main Ensemble
Store
① Copy centres by RR
② Makes predictions on
Projection Set (p′)
Worst from Ensemble (c_worst)
③
④ Get most similar to
Q( )--- Q( )
1
n
⑤ (c_sim)
Yes
Pairwise Q-Statistics (Q_all)
No
Q(,) ≥ θ AND trained with more examples?

Fig. 3. Illustration of the Diversity-based Memory Management Strategy for Class-Imbalanced Learning

Once this is done, base learners in memory are prompted to predict on P' (Step 3 of Fig. 2). Should this strategy is triggered to identify and recover relevant past base learners after b time steps from drift detection, the newly created base learner is also prompted to predict on P'. The prediction correctness of each base learner c_i on P' form a different set of examples for clustering the base learners:

$$C = \{\mathbf{z}_i\}_{i=1}^{L},$$

where $\mathbf{z}_i \in \{0, 1\}^M$ corresponds to the prediction correctness of base learner c_i on P', $M = K_{max} \times 2L$ is the number of projection examples in P', 1 (0) represents a correct (incorrect) prediction, and L is the number of base learners. These examples are called "clustering examples".

Finally, CDCMS.CIL employs an offline clustering method to cluster the model space based on C (Step 4 of Fig. 2). Following the methodology of [11], this work employs Expectation Maximisation (EM) clustering with 10-fold cross-validation to determine the number of clusters. Similar to CDCMS, the clustering result is used either to form a heterogeneously diverse ensemble upon concept drift detection or to identify and recover relevant past base learners after b time steps after drift detection, depending on the context.

E. Diversity-based Memory Management Strategy for Class-Imbalanced Learning

Any diversity-based memory management strategy to decide which base learners to keep in the memory based on the most recent sliding window of examples would suffer similar issues to those of clustering in the model space outlined in Section IV-B. To overcome this, we propose a new diversity-based memory management strategy for class-imbalanced learning, as illustrated in Fig. 3. CDCMS.CIL triggers this strategy when a new base learner is created at regular intervals (b time steps) to replace the least performing one (c_{worst}) in the main ensemble. This strategy determines whether c_{worst} should be stored in the memory when the memory is already full. In particular, Yule's Q-Statistics (Eq. 1) is employed to measure the diversity level between c_{worst} and each past base learner c_i^{memory} in memory (Steps 1-2 of Fig. 3):

$$Q(c_i, c_j) = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}, \quad (1)$$

where $N^{a,b}$ is the number of resampled projection set examples where the classification by c_i (c_j) is a (b), a value of 1 (0) represents a correct (incorrect) classification. Q varies between 1 and -1. Models that tend to classify the same (different) examples correctly will have positive (negative) values of Q.

This strategy uses a smaller projection set p relevant only to classifiers c_{worst} and c_i^{memory}:

$$p = \{\Upsilon^0_{c_{worst}} \cup \Upsilon^1_{c_{worst}} \cup \Upsilon^0_{c_i^{memory}} \cup \Upsilon^1_{c_i^{memory}}\}$$

The resampled projection set p' is

$$p' = \{RR(\Upsilon, k_{max}) | \forall \Upsilon \in p\},$$

where

$$k_{max} = max(|\Upsilon^0_{c_{worst}}|, |\Upsilon^1_{c_{worst}}|, |\Upsilon^0_{c_i^{memory}}|, |\Upsilon^1_{c_i^{memory}}|).$$

Therefore, a vector of Q-Statistics' results is obtained (Step 3 of Fig. 3):

$$Q_{all} = \{Q(c_{worst}, c_1^{memory}), \ldots, Q(c_{worst}, c_n^{memory})\}$$

Next, the most similar past base learner in the memory (c_{sim}) is found by $max(Q_{all})$ (Step 4 of Fig. 3). Finally, we check if $Q(c_{worst}, c_{sim}) \geq \theta$ to determine if they are similar enough (Step 5 of Fig. 3). If they are, the one trained with more examples is retained in the memory (Step 6 of Fig. 3). Otherwise, c_{worst} is discarded and c_{sim} is retained in memory.

V. Experiments

The following approaches were analysed to answer RQ1-3:

- *GH-VFDT$_d$ and HD-VFDT$_d$* [35]: Single-learner baselines using class-imbalance insensitive tree split criteria.
- *Online Bagging$_d$ (OB$_d$)* [36]: A homogeneous diversity ensemble baseline that is not specifically designed to handle concept drift or class-imbalance in data stream learning. The subscript "d" indicates the use of a drift detection wrapper for handling concept drift [18].
- *OOB$_d$ and UOB$_d$* [18]: Simple yet effective homogeneous diversity approaches for class-imbalance data streams, with a drift detection method to enable addressing concept drifts affecting $P(Y)$ and $P(Y|X)$.
- *CSARF* [31]: A state-of-the-art cost-sensitive homogeneous diversity ensemble for drifting class-imbalanced data stream learning.
- *VFC-SMOTE* [30]: A state-of-the-art SMOTE method for drifting class-imbalanced data stream learning. Online Bagging was adopted as its base learner (homogeneous diversity) for fair comparison.
- *SMOTE-OB* [10]: A state-of-the-art homogeneous diversity ensemble integrating random undersampling and VFC-SMOTE's data-level strategy into Online Bagging for drifting class-imbalanced data stream learning.
- *CDCMS.CIL*: Our proposed approach, notable as the only existing heterogeneous diversity approach for class-imbalanced data streams.

A. Data Streams

For practical applicability assessment of heterogeneous and homogeneous diversity, nine real-world streams were used: Airline [37] containing 20 years of commercial flight records; NOAA [38], with five decades of weather measurements; Luxembourg [39], based on European Social Survey data (2002-2007) about internet usage; Ozone [40], comprising air quality data from Houston, Galveston, and Brazoria (1998-2004); PAKDD2009 [41], feature credit scoring data from a major Brazilian retail chain; the Amazon stream [42], with labelled product reviews from 1998 to 2004; the Twitter stream [43], consisting of annotated tweets collected between July and December 2015; Covtype [44] detailing forest cover type of of 30 × 30m cells; INSECTS [45] contains flying data from three insect species, collected via smart traps in a climate controlled setting. Covtype and INSECTS were originally multi-class problems. They have been adapted into several versions of binary classification problems for this study as shown in the supplementary material [46].

Although real-world data streams were used to demonstrate practical performance of the approaches, their concept drifts are unknown, offering limited insight into when and under what circumstances homo/heterogeneous diversity contributes to performance. For a deeper understanding of the benefits and nuances of these diversity approaches, ten artificial data streams from [11] with fully known characteristics were used. Each stream has two variants: abrupt drift (1 time step) and gradual drift (2k time steps). Drift severities and recurrence are detailed in Tables I and II of [11]. We applied four class-imbalance ratios to each variant (0.5:0.5, 0.7:0.3, 0.8:0.2, 0.9:0.1). Overall, 80 artificial streams (10×2×4) were adopted.

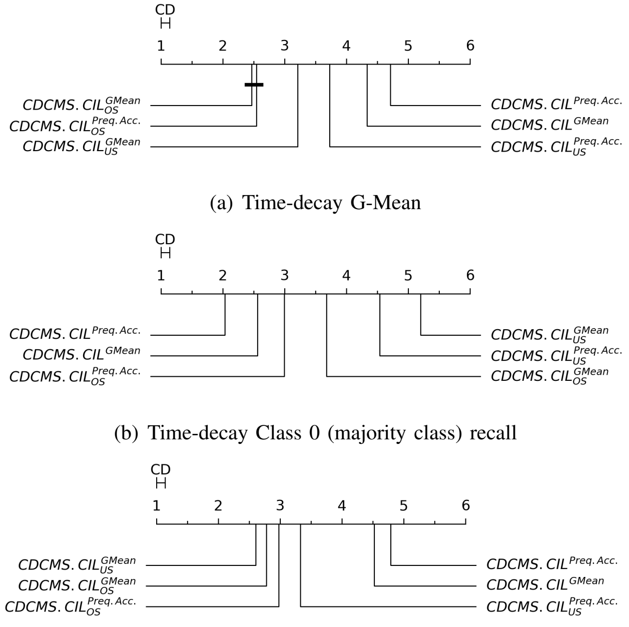
B. Experimental Setup

For the artificial streams (Sine, Agrawal, SEA, STAGGER generators), hyper-parameters were tuned via grid search on four random streams with the hyper-parameter combinations shown in the supplementary material [46]. The configuration with the highest average time-decay G-Mean over ten runs was selected. The same tuning process was applied to the initial 10% of data for real-world streams. A limit of ten micro-clusters per class per base learner was enforced.

The approaches were evaluated by averaging the time-decay G-Mean, class 0 recall and class 1 recall over thirty runs [47]. Time-decay G-Mean was chosen as the primary evaluation metric as it is unbiased for class-imbalanced data [48], capturing both class 1 recall and false-positives (through class 0 recall). Metrics were measured prequentially, sampled every 500 time steps, except for the shorter NOAA and Ozone streams (every 50 steps) and Luxembourg (every 10 steps). A fading factor of 0.999 was used to give more weight to recent performance [47]. Friedman and Nemenyi tests were used to check for statistical significance at the 0.05 level.

C. RQ1: How to Best Leverage Heterogeneity

To address RQ1, we examined the predictive performance of CDCMS.CIL using different weighting metrics in the main



(a) Time-decay G-Mean

(b) Time-decay Class 0 (majority class) recall

(c) Time-decay Class 1 (minority class) recall

Fig. 4. Critical difference of Nemenyi post-hoc test of CDCMS.CIL with different weighting metrics and resampling strategies across all streams. Smaller rank indicates better performance. P-values of Friedman tests are all $\leq 2.2 \times 10^{-16}$. No significant difference was found between approaches linked by horizontal bars.

ensemble with various resampling methods for training. Variations are denoted as "CDCMS.CIL$_{\{\text{resampling method}\}}^{\{\text{weighting metric}\}}$", where "Preq. Acc.", "GMean", "OS", and "US" refer to prequential accuracy, time-decay G-Mean, oversampling, and undersampling, respectively. The absence of 'OS' or 'US' indicates no resampling. This analysis assesses the need for a class-imbalance insensitive weighting metric and resampling to fully exploit heterogeneous diversity within CDCMS.CIL.

The Nemenyi post-hoc critical differences in Fig. 4 show that the weighting metric choice has limited impact on overall predictive performance. Fig. 4(a) shows that, across all data streams, variations using time-decay G-Mean marginally outperformed those using prequential accuracy in time-decay G-Mean ranks, with no significant difference observed when coupled with oversampling. Weighting with time-decay G-Mean generally led to better (worse) minority (majority) class recall compared to prequential accuracy (Figs. 4(c) and 4(b)).

Regarding resampling, oversampling was generally more effective for CDCMS.CIL in terms of time-decay G-Mean (see Fig. 4(a)). CDCMS.CIL$_{\text{OS}}^{\text{Preq. Acc.}}$ and CDCMS.CIL$_{\text{OS}}^{\text{GMean}}$ achieved significantly higher time-decay G-Mean than others, with no significant difference between these two. In contrast, undersampling was not beneficial; CDCMS.CIL$_{\text{US}}^{\text{Preq. Acc.}}$ and CDCMS.CIL$_{\text{US}}^{\text{GMean}}$ had significantly lower time-decay G-Mean, compared to their oversampling counterparts (Figs. 4(c) and 4(b)). This was mainly due to low class 0 recall, despite CDCMS.CIL$_{\text{US}}^{\text{GMean}}$'s high class 1 recall.

Overall, CDCMS.CIL$_{\text{OS}}^{\text{GMean}}$ is typically preferred among the explored variations, achieving the highest G-Mean (alongside CDCMS.CIL$_{\text{OS}}^{\text{Preq. Acc}}$) and only slightly lower minority class recall than the top variation. The top time-decay G-Means achieved by this variation were consistently observed across different imbalance ratios and drift speeds (plots omitted due to space constraints). CDCMS.CIL$_{\text{US}}^{\text{GMean}}$ might be considered if minority class recognition is critical. For highly imbalance ratios (e.g., 0.9:0.1), CDCMS.CIL$_{\text{OS}}^{\text{Preq. Acc}}$ is recommended, as it achieves the highest time-decay G-Mean in this scenario.

TABLE I
FRIEDMAN RANKS OF APPROACHES IN
TIME-DECAY G-MEAN, CLASS 0, AND CLASS 1 RECALLS

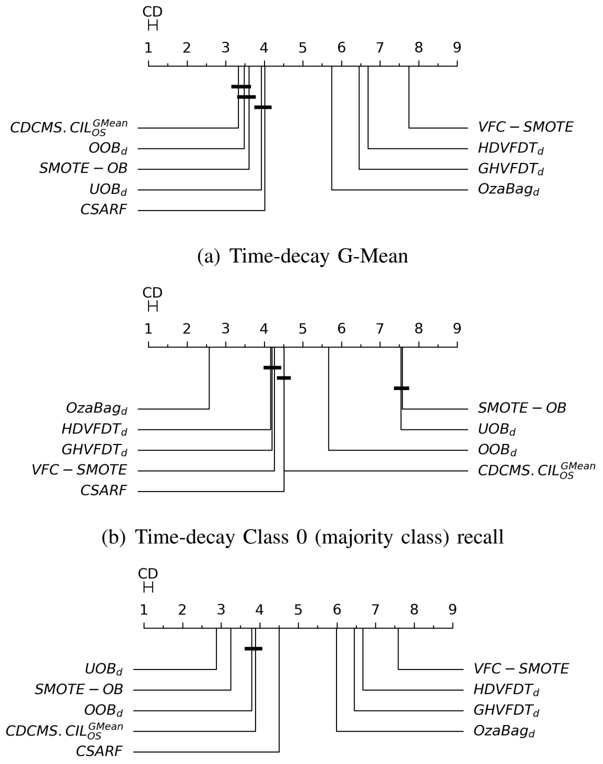
Groups	GH-VFDT$_d$	HD-VFDT$_d$	Oza-Bag$_d$	OOB$_d$	UOB$_d$	CSARF	VFC-SMOTE	SMOTE-OB	CDCMS .CIL$_{OS}^{GMean}$
Time-Decay G-Mean									
Grouped by imbalance ratio (Artificial data streams)									
0.5:0.5	5.90	5.98	3.23	3.99	5.04	4.42	8.16	6.18	2.12
0.7:0.3	6.49	6.65	6.33	3.15	3.84	4.53	7.45	4.11	2.44
0.8:0.2	6.88	7.36	6.84	2.74	2.93	4.77	7.37	2.93	3.18
0.9:0.1	7.25	7.64	6.23	3.19	2.40	4.47	7.26	2.48	4.09
Grouped by concept drift speed (Artificial data streams)									
Abr.	6.49	6.78	5.34	3.23	3.71	4.83	7.44	4.51	2.67
Grad.	6.77	7.04	5.98	3.30	3.39	4.26	7.68	3.34	3.25
Real-world streams (Aggregated)									
Real	5.74	5.79	6.10	4.39	5.51	1.78	8.54	2.24	4.92
Grouped by streams									
Sine	7.99	8.27	6.18	2.46	3.83	2.21	6.40	4.68	2.97
Agr.	5.73	5.67	4.78	4.15	3.83	6.38	8.94	2.52	2.99
SEA	7.57	7.62	7.34	2.64	3.53	4.46	6.72	2.03	3.09
STA.	6.13	7.30	5.21	2.91	2.74	3.30	6.79	7.88	2.75
All	6.46	6.69	5.74	3.48	3.93	4.02	7.75	3.60	3.33
Time-Decay Class 0 (Majority Class) Recall									
Grouped by imbalance ratio (Artificial data streams)									
0.5:0.5	5.82	5.92	3.84	4.77	5.10	3.71	7.73	5.98	2.13
0.7:0.3	3.76	3.49	1.56	6.32	8.42	3.97	4.47	8.15	4.87
0.8:0.2	3.74	3.53	1.78	6.25	8.58	4.31	3.35	8.29	5.18
0.9:0.1	3.30	3.24	2.01	6.05	8.55	4.83	3.15	8.44	5.44
Grouped by concept drift speed (Artificial data streams)									
Abr.	4.29	4.24	2.34	5.82	7.53	3.89	4.95	7.99	3.95
Grad.	4.02	3.85	2.25	5.88	7.79	4.52	4.40	7.44	4.85
Real-world streams (Aggregated)									
Real	4.39	4.67	3.73	4.89	7.03	5.79	2.53	7.00	4.97
All	4.20	4.16	2.57	5.66	7.54	4.51	4.26	7.58	4.51
Time-Decay Class 1 (Minority Class) Recall									
Grouped by imbalance ratio (Artificial data streams)									
0.5:0.5	5.36	5.44	3.33	3.99	4.86	5.01	7.61	6.30	3.11
0.7:0.3	6.72	7.02	6.77	3.59	1.66	5.22	7.43	3.10	3.50
0.8:0.2	6.92	7.43	6.93	3.20	1.64	5.20	7.38	2.60	3.69
0.9:0.1	7.32	7.63	6.37	3.51	1.82	4.74	7.20	2.08	4.33
Grouped by concept drift speed (Artificial data streams)									
Abr.	6.47	6.83	5.53	3.47	2.60	5.32	7.31	4.09	3.39
Grad.	6.69	6.94	6.16	3.67	2.39	4.77	7.50	2.96	3.93
Real-world streams (Aggregated)									
Real	5.90	5.78	6.57	4.73	4.47	2.22	8.36	2.09	4.88
All	6.68	6.89	6.00	3.84	2.29	4.40	7.32	4.92	3.79

- The p-values of Friedman tests are all $\leq 2.2 \times 10^{-16}$.
- Highlighted ranks denote significant superior performance.
- "Abr": Aburpt; "Grad.": Gradual; "Agr.": Agrawal; "STA.": STAGGER;

D. RQ2a&3a: The Benefit of Hetero/Homogeneous Diversity

RQ2a and RQ3a are answered by analysing the predictive performance of the proposed heterogeneous diversity framework (CDCMS.CIL), five existing homogeneous diversity class-imbalanced learning approaches and three baselines as listed at the beginning of Section V. Based on the findings in Section V-C, we focus on CDCMS.CIL$_{\text{OS}}^{\text{GMean}}$.

Fig. 5(a) shows that CDCMS.CIL$_{\text{OS}}^{\text{GMean}}$, which leverages heterogeneous diversity, outperformed non-diversity approaches (GH-VFDT$_\text{d}$ and HD-VFDT$_\text{d}$) and all but one homogeneous ensemble (OOB$_\text{d}$). Although no statistically significant difference was found between the time-decay G-Mean of CDCMS.CIL$_{\text{OS}}^{\text{GMean}}$ and OOB$_\text{d}$ (Fig. 5(a)), CDCMS.CIL$_{\text{OS}}^{\text{GMean}}$



(a) Time-decay G-Mean

(b) Time-decay Class 0 (majority class) recall

(c) Time-decay Class 1 (minority class) recall

Fig. 5. Critical difference of Nemenyi post-hoc test between approaches across all streams. Smaller rank indicates better performance. P-values of Friedman tests are all $\leq 2.2 \times 10^{-16}$. No significant difference was found between approaches linked by horizontal bars.

more consistently achieved higher ranks across a wider range of imbalance ratios and concept drift speeds (see Table I).

Homogeneous diversity was also advantageous. Most class-imbalance-focused homogeneous diversity approaches (OOB$_\text{d}$, UOB$_\text{d}$, CSARF, SMOTE-OB) attained top ranks in at least one factor. Notably, OzaBag$_\text{d}$, a homogeneous approach that was not designed for class-imbalanced learning, outperformed cost-sensitive decision trees (GH-VFDT$_\text{d}$ and HD-VFDT$_\text{d}$) in this context, as shown in Fig. 5(a). These findings emphasise diversity is a crucial factor, possibly more so than cost-sensitive methods in class-imbalanced data stream leaning.

Comparing VFC-SMOTE and SMOTE-OB, both employ the same synthetic minority generation strategy, but VFC-SMOTE generally performed worse than SMOTE-OB and even OzaBag$_\text{d}$. This suggests that applying oversampling independently to each ensemble member (SMOTE-OB) may be more effective for diversity than applying it to the whole ensemble (VFC-SMOTE). Therefore, effectively integrating diversity and class-imbalance strategies is likely more crucial for handling class-imbalance in drifting streams than the sole efficacy of the class-imbalance strategy.

Class-wise performance analysis across all data streams (Figs. 5(b), 5(c)) shows that approaches relying purely on diversity (OzaBag$_\text{d}$), solely on cost-sensitivity (HD-VFDT$_\text{d}$, GH-VFDT$_\text{d}$), or applying class-imbalance strategies less integrally (VFC-SMOTE), performed worse on the minority class (class 1) than the majority (class 0). Conversely, other methods performed better on Class 1. This finding highlights that strategies effectively integrating data-level or cost-sensitive techniques with diversity are better at learning the minority class, thus improving time-decay G-Means.

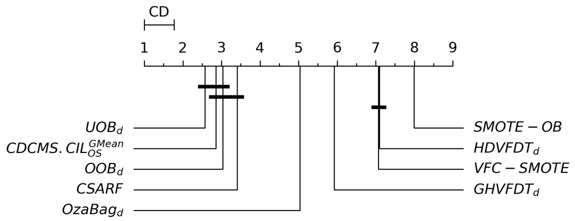


Fig. 6. Critical difference of Nemenyi post-hoc test between approaches in time-decay G-Mean across all STAGGER2 streams. Smaller rank indicates better performance. P-values of Friedman tests are all $\leq 2.2 \times 10^{-16}$. No significant difference was found between approaches linked by horizontal bars.

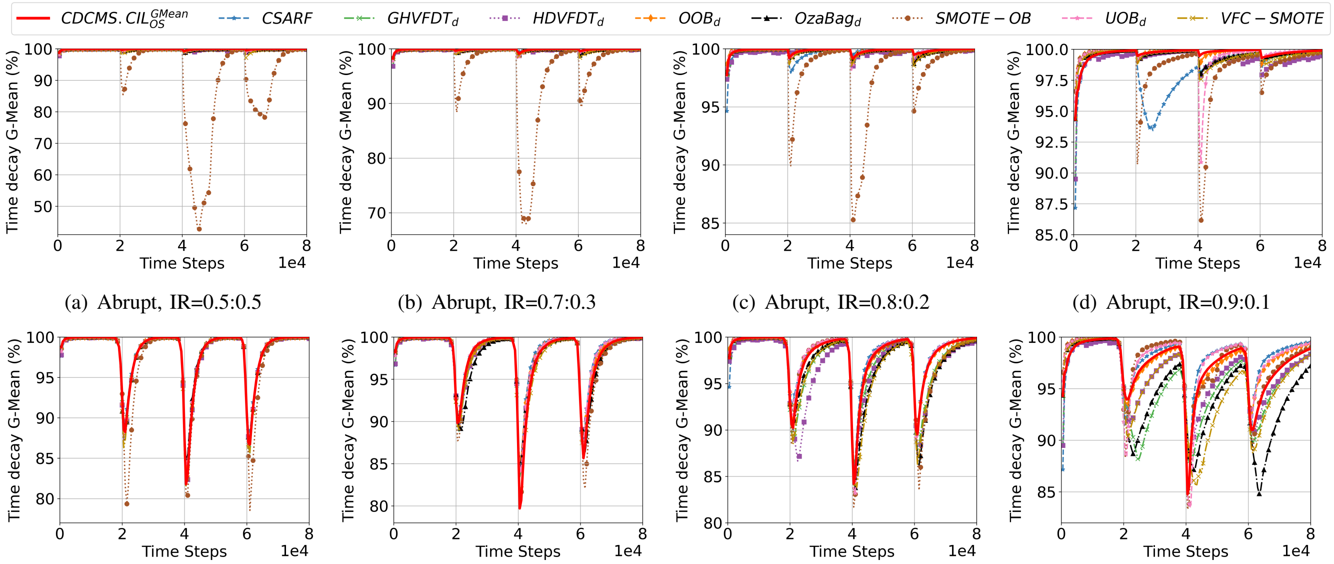
E. RQ2b&3b: When and Why Are Heterogeneous and Homogeneous Diversity Helpful/Detrimental

This section provides a detailed analysis of the predictive performance of the approaches to investigate when heterogeneous and homogeneous diversity are beneficial or detrimental in handling class-imbalanced data streams with concept drift, addressing RQ2b and RQ3b. We use the STAGGER2, SEA2 and PAKDD2009 as representative cases for investigation, as they cover different situations where heterogeneous diversity was more / less helpful than homogeneous approaches. As these situations can also be found on the other data streams, the explanations of the approaches' behaviours on other data streams is similar.

1) STAGGER2 Streams: The benefit of heterogeneous diversity is highlighted in STAGGER2 streams. Fig. 6 shows CDCMS.CIL$_\text{OS}^\text{GMean}$, UOB$_\text{d}$ and OOB$_\text{d}$ as top overall performers across the these streams. Fig. 7 shows CDCMS.CIL$_\text{OS}^\text{GMean}$ maintained consistent performance. In contrast, other approaches showed significant post-drift drops and slow recovery (e.g., SMOTE-OB on all STAGGER2 streams, and CARFF and UOB$_\text{d}$ on IR=0.9:0.1).

Log analysis revealed CDCMS.CIL$_\text{OS}^\text{GMean}$ correctly detected all STAGGER2 concept drifts. For the first and second drifts, it successfully identified them as drifts to new concepts, forming a new main ensemble (NL) from scratch or by reusing relevant past models. This was enabled by its sparse representation and stream clustering (Section IV-C), which captured the spatial distributions of both classes under class-imbalance, allowing accurate retrieval of relevant past models. For the last recurrent drift, the model effectively recognised and recovered models of the first concept, highlighting the clustering strategy's potential in handling class-imbalance (Section IV-D). However, during the severe class-imbalanced gradual drift (STAGGER2-gradual 0.9:0.1), it misidentified the transitional period as an independent concept, recovering models from this phase and achieving only average performance recovery (see Fig. 7(h)).

The heterogeneous diverse NH ensemble, formed by combining models through clustering, mitigated performance drops after abrupt and recurrent drifts, as evident by our analysis



(a) Abrupt, IR=0.5:0.5

(b) Abrupt, IR=0.7:0.3

(c) Abrupt, IR=0.8:0.2

(d) Abrupt, IR=0.9:0.1

(e) Gradual, IR=0.5:0.5

(f) Gradual, IR=0.7:0.3

(g) Gradual, IR=0.8:0.2

(h) Gradual, IR=0.9:0.1

Fig. 7. Time-decay G-Mean of Homo/Heterogeneous Approaches on STAGGER2

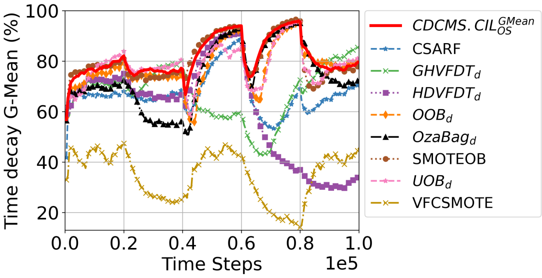


Fig. 8. Performance of Homo/Heterogeneous Approaches on Agrawal3-abrupt 0.8:0.2 (Time-decay G-Mean)

of the logs. In STAGGER2-gradual 0.9:0.1 (Fig. 7(h)), a 0.54 weight to NH reduced post-drift performance decline, demonstrating how heterogeneous diversity supports transitions in imbalanced scenarios, despite sub-optimal recovery due to transitional model retrieval.

OOB$_{\text{d}}$ and UOB$_{\text{d}}$ also ranked highly. Despite lacking heterogeneous diversity, they both performed similarly to CDCMS.CIL$^{\text{GMean}}_{\text{OS}}$. This is likely due to the stream's simplicity, which may not require special strategies to handle drifts. On the complex Agrawal stream (Fig. 8), OOB$_{\text{d}}$ faced significant post-drift performance drops, while CDCMS.CIL$^{\text{GMean}}_{\text{OS}}$ leveraged past models effectively.

2) SEA2 Streams: Homogeneous diversity was more influential in SEA2 streams. CDCMS.CIL$^{\text{GMean}}$OS achieved competitive G-Mean performance with most approaches in SEA2 IR=0.5:0.5, 0.7:0.3, and 0.8:0.2 (Figs. 9 and 10). However, in severely imbalanced SEA2 streams (0.9:0.1), it was outperformed by UOB$_{\text{d}}$ and SMOTE-OB, though it performed similarly to CSARF and OOB$_{\text{d}}$.

Analysis of CDCMS.CIL$^{\text{GMean}}$OS logs revealed minimal concept drift detection in SEA2 streams, likely due to the low severity of the SEA drifts. Therefore, the activation and impact of CDCMS.CIL's heterogeneous diversity strategy was limited. Even when activated, its effect was minimal due to the low drift severity. CSARF may have faced a similar issue of drift detection, hindering its drift adaptation mechanisms.

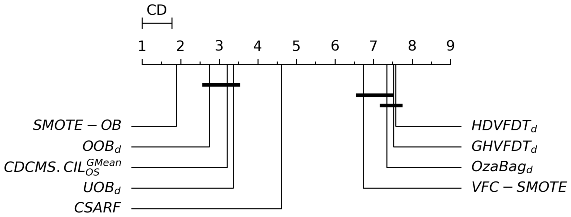
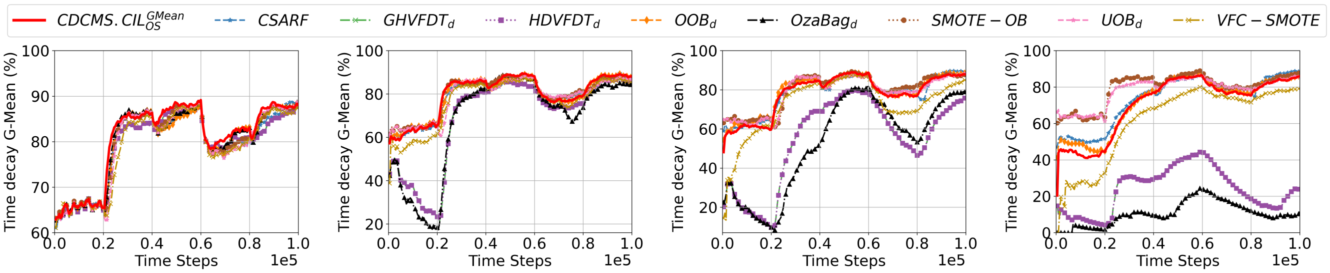


Fig. 9. Critical difference of Nemenyi post-hoc test between approaches in time-decay G-Mean across all SEA2 streams. Smaller rank indicates better performance. P-values of Friedman tests are all $\leq 2.2\times10^{-16}$. No significant difference was found between approaches linked by horizontal bars.

The infrequent use of heterogeneous diversity means performance relied primarily on applying class-imbalance learning strategies within a homogeneously diverse ensemble. These strategies can be implemented at two levels: the ensemble level, where the training data is balanced for the ensemble as a whole; and the base learner level, where the balance is adjusted individually for each constituent base learner, potentially allowing for deeper integration with diversity. Figs. 10(c) and 10(d) shows that VFC-SMOTE, which applies a class-imbalanced strategy at the ensemble level, was less effective in SEA2 IR=0.8:2 and 0.9:1, indicating class-imbalance strategies might be better applied at the base learner level through methods like cost-sensitive learning or data subsetting, which introduce diversity. SMOTE-OB and UOB$_{\text{d}}$ performed well in SEA IR=0.9:0.1, likely benefiting from undersampling for severely imbalanced data streams [18]. SMOTE-OB slightly outperformed UOB$_{\text{d}}$, which is attributed to its use of both oversampling and undersampling at the base learner level, a strategy that tends to yield better diversity.

A related scenario is the Luxemburg stream, where stable time-decay G-Means across the stream suggests the absence of concept drifts. This favoured approaches that do not rely



(a) Abrupt, IR=0.5:0.5 (b) Abrupt, IR=0.7:0.3 (c) Abrupt, IR=0.8:0.2 (d) Abrupt, IR=0.9:0.1

Fig. 10. Time-decay G-Mean of Homo/Heterogeneous Approaches on SEA2 abrupt drifts (the results were very similar for gradual drifts)

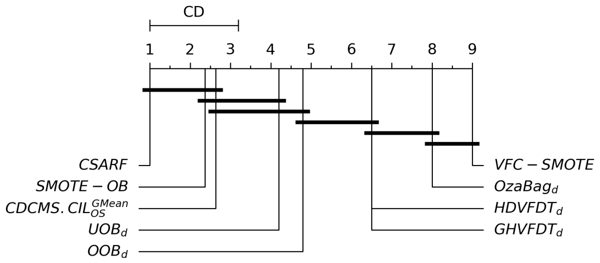
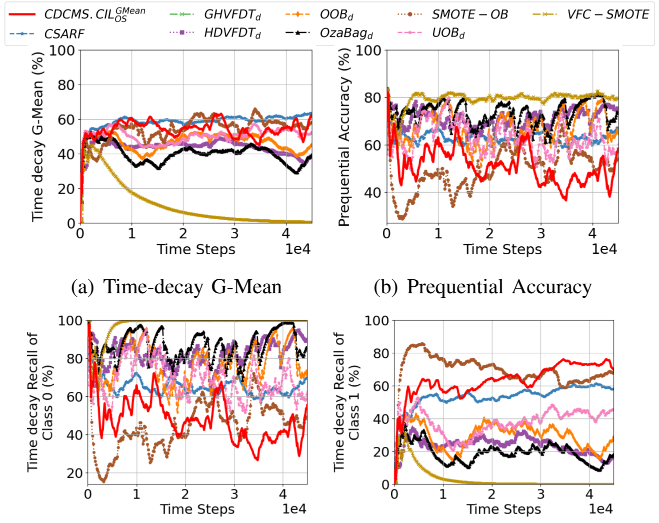


Fig. 11. Critical difference of Nemenyi post-hoc test between approaches in time-decay G-Mean on PAKDD2009. Smaller rank indicates better performance. P-value of Friedman test $\leq 2.2 \times 10^{-16}$. No significant difference was found between approaches linked by horizontal bars.



(a) Time-decay G-Mean (b) Prequential Accuracy

(c) Time-decay Recall in Class 0 (d) Time-decay Recall in Class 1

Fig. 12. Time-Decay G-Mean of Homo/Heterogeneous Approaches on PAKDD2009

on drift detection, and was detrimental to approaches that do, such as CDCMS.CIL, which can be negatively affected by false positive drift detections.

3) PAKDD2009: Fig. 11 shows that CSARF, $\text{CDCMS.CIL}_{\text{OS}}^{\text{GMean}}$, and SMOTE-OB performed the best on time-decay G-Mean in PAKDD2009, with no significant difference among them. Fig. 12(a) further shows that CSARF maintained consistently high time-decay G-Mean throughout the stream. $\text{CDCMS.CIL}_{\text{OS}}^{\text{GMean}}$ performed in a range similar to SMOTE-OB, often surpassing it before the 30k-th time step. Both occasionally achieved performance comparable to CSARF. Most other approaches struggled to perform well.

Figs. 12(c) and 12(d) provide the class-wise details. Majority class performance (Fig. 12(c)) fluctuated for most approaches, suggesting potential concept drifts or unstable class-imbalanced learning strategies. Minority class performance (Fig. 12(d)) was notably better for $\text{CDCMS.CIL}_{\text{OS}}^{\text{GMean}}$, CSARF, and SMOTE-OB, which have deeper integration of diversity and class-imbalance strategies. Notably, $\text{CDCMS.CIL}_{\text{OS}}^{\text{GMean}}$ gradually improved its minority class recall throughout the stream, rising from around 60% to nearly 80%. In contrast, SMOTE-OB exhibited a reverse trend, with minority class recall dropping from over 80% to as low as 60%. CSARF showed a modest improvement in minority class recall, briefly catching up with SMOTE-OB around the 35k-th time step. These results highlight $\text{CDCMS.CIL}_{\text{OS}}^{\text{GMean}}$'s efficacy in enhancing minority class performance in PAKDD2009 and, again, imply that heterogeneous diversity may be particularly beneficial for minority class performance during drifts.

VI. Conclusion

This study investigated the impact of homogeneous and heterogeneous diversity in drifting class-imbalanced data streams and introduced CDCMS.CIL, a heterogeneous diversity ensemble. Experiments revealed that CDCMS.CIL best leveraged heterogeneous diversity for optimal performance when using time-decay G-Mean weighting combined with oversampling, but suggest undersampling for applications requiring accurate minority class identification, and time-decay prequential accuracy with oversampling for extreme class-imbalance (RQ1). Heterogeneous diversity in CDCMS.CIL significantly improved adaptation to recurring and severe concept drifts by leveraging past models, often surpassing other approaches. However, its effectiveness was hindered by false-positive drift detections and mistaking gradual drifts for past concepts, especially under severe imbalance. In contrast, homogeneous diversity was effective during stable concepts or mild drifts (RQ2). Our findings indicate that role of diversity in handling drifting class-imbalanced data streams may be even more critical than that of class-imbalance strategies, and that a deep integration of diversity and class-imbalance strategies is crucial to achieve top performance. For example, applying class-imbalance strategies at the base learner level enhanced performance by promoting greater diversity (RQ3).

Future work will focus on reducing CDCMS.CIL's dependency on drift detection, improving its robustness, and analysing computational complexity. The use of a wide range of datasets suggests real-world applicability, but further analyses with more data streams would enhance the generalisability of the conclusions.

References

[1] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A Survey on Concept Drift Adaptation," *ACM Computing Survey*, vol. 46, no. 4, 03 2014.

[2] G. Aguiar, B. Krawczyk, and A. Cano, "A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework," *Mach. Learn.*, 2023.

[3] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit card fraud detection: A realistic modelling and a novel learning strategy," *IEEE TNNLS*, pp. 1–14, 09 2017.

[4] Kanika, J. Singla, A. K. Bashir, Y. Nam, N. U. Hasan, and U. Tariq, "Handling class imbalance in online transaction fraud detection," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 2861–2877, 2022.

[5] F. Jemili, K. Jouini, and O. Korbaa, "Intrusion detection based on concept drift detection and online incremental learning," *IJPCC*, vol. 21, no. 1, pp. 81–115, Jan 2025.

[6] A. A. Toor, M. Usman, F. Younas, A. C. M. Fong, S. A. Khan, and S. Fong, "Mining massive e-health data streams for iomt enabled healthcare systems," *Sensors*, vol. 20, no. 7, 2020.

[7] S. Wang, L. L. Minku, and X. Yao, "A Systematic Study of Online Class Imbalance Learning With Concept Drift," *IEEE TNNLS*, vol. 29, no. 10, pp. 4802–4821, 2018.

[8] ——, "Resampling-Based Ensemble Methods for Online Class Imbalance Learning," *IEEE TKDE*, vol. 27, no. 5, pp. 1356–1368, 2015.

[9] B. Wang and J. Pineau, "Online bagging and boosting for imbalanced data streams," *IEEE TKDE*, vol. 28, no. 12, pp. 3353–3366, 2016.

[10] A. Bernardo and E. D. Valle, "SMOTE-OB: Combining SMOTE and Online Bagging for Continuous Rebalancing of Evolving Data Streams," in *BigData*, 2021, pp. 5033–5042.

[11] C. W. Chiu and L. L. Minku, "A Diversity Framework for Dealing With Multiple Types of Concept Drift Based on Clustering in the Model Space," *IEEE TNNLS*, vol. 33, no. 3, pp. 1299–1309, 2022.

[12] G. Ditzler and R. Polikar, "Incremental learning of concept drift from streaming imbalanced data," *IEEE TKDE*, vol. 25, no. 10, pp. 2283–2301, 2013.

[13] C. W. Chiu and L. L. Minku, "Diversity-Based Pool of Models for Dealing with Recurring Concepts," in *IJCNN*, 07 2018, pp. 2759–2766.

[14] L. L. Minku, *Transfer Learning in Non-stationary Environments: Methods and Applications*, 01 2019, pp. 13–37.

[15] I. Žliobaitė, "Learning under concept drift: an overview," *CoRR*, vol. abs/1010.4784, 01 2010.

[16] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar, "Learning in nonstationary environments: A survey," *IEEE CIM*, vol. 10, pp. 12–25, 11 2015.

[17] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak, "Ensemble learning for data stream analysis: A survey," *Information Fusion*, vol. 37, p. 132–156, 09 2017.

[18] C. W. Chiu and L. L. Minku, "SMOClust: Synthetic Minority Oversampling based on Stream Clustering for Evolving Data Streams," *Mach. Learn.*, vol. 113, p. 4671–4721, 2024.

[19] L. L. Minku, A. P. White, and X. Yao, "The Impact of Diversity on Online Ensemble Learning in the Presence of Concept Drift," *IEEE TKDE*, vol. 22, no. 5, pp. 730–742, 2010.

[20] L. L. Minku and X. Yao, "DDD: A New Ensemble Approach for Dealing with Concept Drift," *IEEE TKDE*, vol. 24, pp. 619–633, 05 2012.

[21] T. Museba, F. V. Nelwamondo, and K. Ouahada, "Ades: A new ensemble diversity-based approach for handling concept drift," *Mob. Inf. Syst.*, pp. 5 549 300:1–5 549 300:17, 2021.

[22] S. Yin, G. Liu, Z. Li, C. Yan, and C. Jiang, "An accuracy-and-diversity-based ensemble method for concept drift and its application in fraud detection," in *ICDMW*, 2020, pp. 875–882.

[23] Y. Sun, K. Tang, Z. Zhu, and X. Yao, "Concept drift adaptation by exploiting historical knowledge," *IEEE TNNLS*, vol. 29, no. 10, pp. 4822–4832, 2018.

[24] O. A. Mahdi, E. Pardede, N. Ali, and J. Cao, "Diversity measure as a new drift detection method in data streaming," *Knowledge-Based Systems*, vol. 191, p. 105227, 2020.

[25] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," *Artificial Intelligence Review*, vol. 57, no. 6, p. 137, May 2024.

[26] Z. Yan, H. Dong, G. Kang, L. Zhang, and Y.-C. Chen, "Dynamic weighted selective ensemble learning algorithm for imbalanced data streams," *J. Supercomput.*, vol. 78, no. 9, pp. 11 823–11 848, 2022.

[27] H. Zhu, M. Zhou, G. Liu, Y. Xie, S. Liu, and C. Guo, "Nus: noisy-sample-removed undersampling scheme for imbalanced classification and application to credit card fraud detection," *IEEE TCSS*, vol. 10, no. 5, pp. 2402–2412, 2023.

[28] A. Bernardo, H. M. Gomes, J. Montiel, B. Pfahringer, A. Bifet, and E. D. Valle, "C-SMOTE: Continuous Synthetic Minority Oversampling for Evolving Data Streams," 2020, pp. 483–492.

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, p. 321–357, 6 2002.

[30] A. Bernardo and E. Della Valle, "VFC-SMOTE: very fast continuous synthetic minority oversampling for evolving data streams," *Data Mining and Knowledge Discovery*, vol. 35, 11 2021.

[31] Loezer, Lucas and Enembreck, Fabrício and Barddal, Jean Paul and de Souza Britto, Alceu, "Cost-Sensitive Learning for Imbalanced Data Streams," in *SAC*, ser. SAC '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 498–504.

[32] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," *IEEE VLDB*, p. 81–92, 2003.

[33] J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. P. L. F. d. Carvalho, and J. a. Gama, "Data stream clustering: A survey," *ACM Computing Surveys*, vol. 46, no. 1, 10 2013.

[34] R. Moulton, H. Viktor, N. Japkowicz, and J. Gama, *Clustering in the Presence of Concept Drift*, 01 2019, pp. 339–355.

[35] R. Lyon, J. Brooke, J. Knowles, and B. Stappers, "Hellinger Distance Trees for Imbalanced Streams," in *ICPR*, 05 2014.

[36] N. C. Oza, "Online bagging and boosting," in *SMC*, vol. 3, 2005, pp. 2340–2345.

[37] E. Ikonomovska, J. a. Gama, and S. Džeroski, "Learning model trees from evolving data streams," *Data Mining and Knowledge Discovery*, vol. 23, pp. 128–168, 07 2011.

[38] R. Elwell and R. Polikar, "Incremental Learning of Concept Drift in Nonstationary Environments," *IEEE TNN*, vol. 22, no. 10, pp. 1517–1531, 2011.

[39] I. Žliobaitė, "Combining similarity in time and space for training set formation under concept drift," *Intelligent Data Analysis*, vol. 15, pp. 589–611, 06 2011.

[40] K. Zhang, W. Fan, X. Yuan, I. Davidson, and X. Li, "Forecasting Skewed Biased Stochastic Ozone Days: Analyses and Solutions," vol. 14, 12 2006, pp. 753–764.

[41] T. Theeramunkong, B. Kijsirikul, N. Cercone, and T. B. Ho, Eds., *Advances in Knowledge Discovery and Data Mining, 13th Pacific-Asia Conference, PAKDD 2009, Bangkok, Thailand, April 27-30, 2009, Proceedings*, ser. Lecture Notes in Computer Science, vol. 5476. Springer, 2009.

[42] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Annual Meeting of the ACL*. Association for Computational Linguistics, Jun. 2007, pp. 440–447.

[43] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "Semeval-2016 task 4: Sentiment analysis in twitter," *CoRR*, vol. abs/1912.01973, 2019.

[44] J. A. Blackard and D. J. Dean, "Comparative Accuracies of Artificial Neural Networks and Discriminant Analysis in Predicting Forest Cover Types from Cartographic Variables," *Computers and Electronics in Agriculture*, vol. 24, pp. 131–151, 12 1999.

[45] V. M. A. Souza, D. M. dos Reis, A. G. Maletzke, and G. E. A. P. A. Batista, "Challenges in Benchmarking Stream Learning Algorithms with Real-world Data," *Data Mining and Knowledge Discovery*, vol. 34, pp. 1805–1858, 2020.

[46] C. W. Chiu and L. L. Minku, "The value of diversity for dealing with concept drift in class-imbalanced data streams – supplementary document," 2025. [Online]. Available: https://github.com/michaelchiucw/CDCMS.CIL/blob/main/The_Value_of_Diversity_for_Dealing_with_Concept_Drift_in_Class_Imbalanced_Data_Streams_Supplementary_Document.pdf

[47] J. Gama, R. Sebastião, and P. P. Rodrigues, "On evaluating stream learning algorithms," *Mach. Learn.*, vol. 90, pp. 317–346, 10 2013.

[48] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Patt. Recogn.*, vol. 91, pp. 216–231, 2019.