Exploring the Open World Using Incremental Extreme Value Machines

Tobias Koch*, Felix Liebezeit*, Christian Riess†, Vincent Christlein†, and Thomas Köhler*
*e.solutions GmbH, Erlangen, Germany
Email: tobias.koch@esolutions.de
†Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

***Abstract*—Dynamic environments require adaptive applications. One particular machine learning problem in dynamic environments is open world recognition. It characterizes a continuously changing domain where only some classes are seen in one batch of the training data and such batches can only be learned incrementally. Open world recognition is a demanding task that is, to the best of our knowledge, addressed by only a few methods. This work introduces a modification of the widely known Extreme Value Machine (EVM) to enable open world recognition. Our proposed method extends the EVM with a partial model fitting function by neglecting unaffected space during an update. This reduces the training time by a factor of 28. In addition, we provide a modified model reduction using weighted maximum K-set cover to strictly bound the model complexity and reduce the computational effort by a factor of 3.5 from $2.1\,\mathrm{s}$ to $0.6\,\mathrm{s}$. In our experiments, we rigorously evaluate *openness* with two novel evaluation protocols. The proposed method achieves superior accuracy of about $12\,\%$ and computational efficiency in the tasks of image classification and face recognition.**

***Index Terms*—classification and clustering, online learning and continual learning**

I. Introduction

Traditionally, machine learning treats the world as *closed* and *static* space. In particular for classification, domain data is assumed to comprise pre-defined classes with stationary class-conditional distributions. Also datasets to fit models before deploying them shall be available in a single chunk. Practitioners develop such models under controlled lab conditions, where they nowadays rely on tremendous computational resources.

This scarcely applies to many real-world applications as the world is an *open* space in many facets. For instance, classifiers might be confronted with classes unseen during training. Also distributions of pre-trained classes might be non-stationary or models shall learn novel classes within operation mode. These aspects often occur simultaneously like in image classification, where unknown image categories should be distinguished from known ones showing *concept drifts* (*e. g.*, captured new data with different cameras). It is also in the very nature of biometric systems like face or writer identification that are confronted with known subjects having concept drifts (*e. g.*, due to aging or environmental changes), novel subjects to enroll, and unknown subjects. There is also a steady quest for making the respective algorithms computationally efficient to be applicable on edge devices with limited resources.

Open world recognition (OWR) as formalized by Bendale and Boult [1] addresses such constraints and includes three subtasks. 1) *Recognize* new samples either as a *known* or *unknown*. 2) *Label* new samples either by approving the recognition or defining a new known class. 3) *Adapt* the current model by exploiting updated labels.

The recognition subtask poses an independent research area termed open set recognition (OSR) [2] and received a lot of interest in applications like face recognition [3], novelty and intrusion detection [4]–[6], and forensics [7]–[9]. Currently Extreme Value Machine (EVM) models as proposed by Rudd *et al*. [10] are state of the art in OSR. EVMs predict unnormalized class-wise probabilities for query samples to be included in the respective known classes. Model fitting depends on class negatives, *i. e.*, it adapts well to imbalanced data, which is a common problem in incremental learning [11], [12]. However, fitting and prediction scale badly for large datasets making their use on resource limited platforms difficult.

Model adaptability can be achieved by cyclic retraining. However, this model-agnostic approach is computationally inefficient and all data needs to be organized in a single chunk. *Incremental learning* aims at doing adaptions effectively and efficiently by batch-wise or sample-wise incorporation of novel data. This needs to handle different challenges: On the one hand, data undergoes concept drifts that shall be learned. On the other hand, the stability-plasticity dilemma [13] could either lead to maximum predictive power on previously learned classes (*i. e.*, high stability) or on novel classes (*i. e.*, high plasticity). A good tradeoff between both border cases is desired for well-generalizing models. Although there are several incremental formulations of popular classifiers [14], [15] or deep learning architectures [12], [16], [17], these approaches assume closed sets of known classes in their prediction phase. In principle, probabilistic models like the EVM can handle batch-wise data but their actual behaviour in incremental learning under an open world regime is still widely unexplored. In this paper, we show that simple ad-hoc applications of existing EVM approaches in OWR lead to suboptimal stability-plasticity tradeoffs.

The contribution of this paper can be summarized as follows: 1) A partial model fitting algorithm that prevents costly Weibull estimations by neglecting unaffected space during an update. This reduces the incremental training time by a factor of 28. 2) A model reduction technique using weighted maximum K-set cover providing fixed size model complexities, which is fundamental for memory constrained systems. This approach is up to $4\times$ faster than existing methods and achieves higher

Accepted at ICPR 2022.

recognition rates of about $12\,\%$. 3) Two novel open world protocols that can be adapted to vary the task complexity in terms of openness. 4) The framework is evaluated on these protocols with varying difficulty and dimensional complexity for applications such as image classification and face recognition.

II. Related Work

1) Incremental Learning: Popular classifiers such as Support Vector Machines (SVMs), decision trees, linear discriminant analysis, and ensemble techniques are modified to allow efficient model adaptations [14], [15], [18]–[20]. Curriculum and self-paced learning are concepts to sequentially incorporate samples into a model in a meaningful order [21]–[23]. iCaRL [16] and EEIL [17] use distillation or bias correction [12] to counter catastrophic forgetting. Zhang *et al.* [24] proposed a pseudo incremental learning paradigm by decoupling the feature and classification learning stages. However, the adaptation of underlying deep neural networks (DNNs) on embedded hardware, as required in many open world applications [1], is far from being efficient. Additionally, these incremental strategies are not designed for OSR.

2) Open Set Recognition: Early approaches [25]–[28] define threshold-based unknown detection rules for closed-set classifier outputs. More recent methods focus on the Extreme Value Theory (EVT) to consider negative class samples for the estimation of rejection probabilities. Scheirer *et al.* [29] developed the Weibull SVM (W-SVM) that combines a one-class and a binary SVM, where decision scores are calibrated via Weibull distributions. Jain *et al.* [30] proposed the Probability of Inclusion SVM (P_I-SVM) to calibrate the outputs of a RBF SVM to unnormalized posterior probabilities. The related OpenMax [4] calibration is used for class activations of DNNs to model the probability of samples being unknown. Unfortunately, such re-calibrations do not support incremental learning off-the-shelf. Also GANs allow to sharpen open set models with adversarial samples [31]–[34]. Recent novelty detection approaches focus on the uncertainty expressiveness of classifiers that can be used to perform novelty or unknown detection, such as Bayesian neural networks [35], Bayesian logistic regression [7], and Gaussian processes [9]. While these methods commonly require multiple computationally demanding Monte Carlo draws to calculate the predictive uncertainty, Sun *et al.* [36] propose a non-incremental post hoc approach to handle model overconfidence.

3) Open World Recognition: Nearest Neighbor (NN) based classifiers are open world capable, as they typically have no actual training step. The Open Set NN (OSNN) [37] defines the open space via a threshold on the ratio of similarity scores of the two most similar classes. Bendale and Boult [1] derived the Nearest Non-Outlier (NNO) algorithm from the Nearest Class Mean (NCM) classifier [38], [39]. NNO rejects samples that are not in the range of any class center where the distance depends on a learned Mahalanobis distance. However, these approaches are purely distance-based and do not take distributional information into account. Joseph *et al.* [40] proposed an open world object detection method that includes fine-tuning of a DNN which is typically too costly for embedded hardware. To overcome the limitations of NNs, Rudd *et al.* [10] introduced the EVM that defines sample-wise inclusion probabilities in dependence of their neighborhood of other classes. Since this approach is based on a NN-like data structure, they propose a model reduction technique to keep the most relevant data points, similar to the support vectors of SVMs, to reduce the memory footprint. The EVM has achieved state-of-the-art results in intrusion detection [5] and open set face recognition [3]. The C-EVM [41] performs a clustering prior to the actual EVM fitting to reduce the dataset size. These centroids are then used to fit the EVM. However, the clustering does not ensure a reduced model size and especially for small batches, it can cause computational overhead. In contrast, our proposed method adequately detects unaffected space in incremental updates and prevents redundant parameter estimations. Additionally, we provide a computationally more efficient model reduction using weighted maximum K-set cover, that reduces the model size to a fixed user-set value.

III. Background: Extreme Value Theory

The EVM estimates per-sample probabilities of inclusions. Let $\boldsymbol{x}_i$ be a feature vector of class y_i referred to as an anchor sample. Given $(\boldsymbol{x}_i, y_i)$, we select the τ nearest negative neighbors $\boldsymbol{x}_j$, $j = 1, \ldots, \tau$ from different classes $y_j \neq y_i$ according to a distance $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$, where τ denotes a tail size. The inclusion probability of a sample $\boldsymbol{x}$ for class y_i is given by the cumulative Weibull distribution:

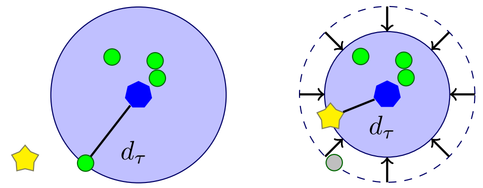
$$\Psi_i(\boldsymbol{x}) = \Psi(\boldsymbol{x}; \theta_i) = \exp\left(-\left(\frac{d(\boldsymbol{x}_i, \boldsymbol{x})}{\lambda_i}\right)^{\kappa_i}\right) \quad , \tag{1}$$

where $\theta_i = \{\kappa_i, \lambda_i\}$ denotes the Weibull parameters, κ_i is the *shape*, and λ_i is the *scale* associated with $\boldsymbol{x}_i$. Given labeled training data $\mathcal{N} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$, each feature vector $\boldsymbol{x}_i$ with class label y_i becomes an anchor. Fitting the underlying EVM aims at sample-wise estimating their θ. A query sample $\boldsymbol{x}$ is assigned to class y_i with maximum probability $\max_{i \in N} \Psi_i(\boldsymbol{x})$. This probability shall reach a threshold δ to distinguish knowns and unknowns according to:

$$y = \begin{cases} y_i & \text{if } \max_{i \in N} \Psi_i(\boldsymbol{x}) \geq \delta \ , \\ \text{“unknown”} & \text{otherwise } . \end{cases} \tag{2}$$

A baseline approach keeps all θ_i, which is expensive in terms of prediction time complexity and memory footprint. Rudd *et al.* [10] proposed a model reduction such that only informative θ_i, *extreme vectors (EVs)*, are kept since samples within the same class might be redundant. It can be expressed as set cover problem [42] to find a minimum number of samples that *cover* all other samples. Redundancies are determined by inclusion probabilities $\Psi_i(\boldsymbol{x}_j)$ within N_c samples of a class c ($y_i = y_j \, \forall i, j \in \{1, \ldots, N_c\}$). A sample $\boldsymbol{x}_j$ is discarded if it is covered by θ_i, *i. e.*, $\Psi_i(\boldsymbol{x}_j) \geq \zeta$, where ζ denotes the coverage threshold. This can be formulated as the minimization problem:

$$\text{minimize} \ \sum_{i=1}^{N_c} I(\theta_i) \ \text{ subject to } \ I(\theta_i)\Psi_i(\boldsymbol{x}_j) \geq \zeta \quad , \tag{3}$$



(a) No update required. (b) Update required.

Fig. 1. Incremental update illustration with $\tau = 4$. The Weibull distribution of the extreme vector (EV) (●) is estimated on the τ nearest samples (●). The blueish hypersphere with radius d_τ is derived from the farthest sample. The new sample (☆) in Figure 1a lies outside the sphere and can be ignored. Once a new sample lies within the sphere, *cf*. Figure 1b, an update is required.

where the indicator function $I(\theta_i)$ is given by:

$$I(\theta_i) = \begin{cases} 1 & \text{if any } \Psi_i(\boldsymbol{x}_j) \geq \zeta \quad \forall j \in N_c \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{4}$$

Rudd *et al.* [10] determines approximate solutions in $\mathcal{O}(N_c^2)$ using greedy iterations, where in each iteration samples that cover most other samples are selected. This approach does not constrain the amount of EVs, which might be necessary for memory limited systems. To this end, bisection to determine a suitable ζ *per class* can be performed.

IV. Incremental Extreme Value Learning

During online learning new data points arise and may interfere with the current EVs' Weibull distribution estimates.

1) Incremental Learning Framework: EVM learning involves two subtasks: 1) *Model fitting* to adapt the model to new data and 2) *model reduction* that bounds the model's computational complexity and required resources. In OWR, both steps need to handle training data arriving batch-wise over consecutive epochs. We perform incremental learning over epochs using new arriving training batches $\mathcal{N}^t$, where t denotes the epoch index. For an incremental formulation, let $\Theta_E^t = \{\theta_1^t, \ldots, \theta_E^t\}$ be a model of E EVs determined either at the previous epoch or learned from scratch at the first epoch. The fit function incorporates the new batch $\mathcal{N}^t$ to the current model Θ_E^t to obtain a new intermediate model Θ^{t+1}. The reduction squashes Θ^{t+1} according to a given budget by selecting most informative EVs considering both previous and new samples. This yields the consolidated model $\Theta_E^{t+1} \subseteq \Theta^{t+1}$. Our framework alternates the fit and reduction function efficiently per epoch.

2) Partial Model Fitting: For model fitting, we process samples in new arriving batches $\mathcal{N}^t$ independently to incorporate them into the current model Θ_E^t. A new sample $\boldsymbol{x}^{t+1}$ might fall into the neighborhood of any EV's feature vector $\boldsymbol{x}_e^t$, which would invalidate the corresponding Weibull parameters in θ_e^t, where $\theta_e^t \in \Theta_E^t$. A naive approach is to re-estimate a new Weibull distribution for each EV including nearest negative neighbor search and tail construction. We argue that this is highly inefficient since it is most likely that the new sample will not influence all the EVs. Thus, most estimates will result in the same Weibull parameters as previously.

TABLE I

Update ratio [%] of the extreme vectors (EVs) on a subset of MNIST. The lower the ratio the more updates can be skipped.

Batch Size	**Tail Size τ** 5	25	100	250
5	0.56	2.44	8.93	20.17
25	2.68	10.81	33.29	59.83
50	5.15	19.26	51.55	79.40
100	9.63	32.54	72.22	93.42
250	21.19	58.05	93.04	99.51

We extend the EVM model by an automatically derivable, *i. e.*, nonuser-set value, namely the *maximum tail distance* d_τ, which corresponds to the maximum distance within a tail such that $\theta_e^t = \{\kappa_e^t, \lambda_e^t, d_{\tau,e}^t\}$. This parameter operates as a threshold and controls the model update. It can be described by a hypersphere centered at an EV with radius d_τ as depicted in Figure 1. Anytime a sample falls into this hypersphere, we need to shrink it. To perform partial fits, we need to compute distances between $\boldsymbol{x}^{t+1}$ and all $\boldsymbol{x}_e^t$ and estimate the Weibull parameters for $\boldsymbol{x}^{t+1}$. Using these distances, we define the update rule for the EV:

$$\theta_e^{t+1} = \begin{cases} \text{update}(\theta_e^t) & \text{if } d(\boldsymbol{x}_e^t, \boldsymbol{x}^{t+1}) < d_{\tau,e}^t \ , \\ \theta_e^t & \text{otherwise} \ , \end{cases} \tag{5}$$

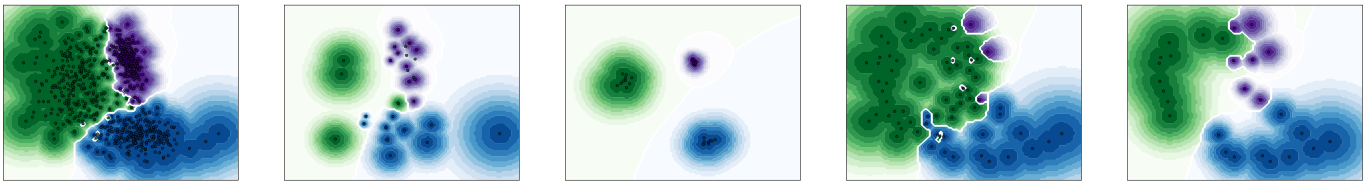
where update$(\cdot)$ denotes tail update, re-estimation of Weibull parameters, and storage of new maximum tail distances. This allows computationally efficient partial fits and leads to exactly the same result as cyclic retraining, as long as no model reduction is carried out.

In Table I, we exemplify the gain of this approach. We incrementally fit an EVM on a subset of MNIST and store all samples as EVs. The update ratio determines the fraction of EVs that require an update in subsequent epochs. It follows, the smaller the batches and tail size the less updates are necessary. The benefit can become very substantial at small batch and tail sizes with an update ratio of only 0.56 %.

3) Model Reduction: In our incremental learning framework, the aim of a class-wise model reduction g is to find a subset $\Theta_{E_c}^t \subseteq \Theta_c^t$ that is budgeted w. r. t. the number of resulting EVs.

a) Problem Statement: For the sake of simplicity, let us drop the batch count t and class index c, unless it is necessary. We denote our model reduction by a function $g\colon \Theta \to \Theta_E$, where Θ_E underlies the constraint $|\Theta_E| \leq K \leq |\Theta| = N$ and K denotes the budget of EVs that can be kept for a certain class with N samples. The intuition behind the design of g is three-fold: 1) We aim at selecting EVs that best cover others according to pair-wise inclusion probabilities. 2) While pair-wise inclusion probabilities are not symmetric in general, *i. e.*, $\Psi_i(\boldsymbol{x}_j) \neq \Psi_j(\boldsymbol{x}_i)$, high bilateral coverage is common and would introduce a bias towards selecting EVs very close to class centroids implying that selecting both $\Psi_i(\boldsymbol{x}_j)$ and $\Psi_j(\boldsymbol{x}_i)$ shall be penalized. 3) At most K EVs shall be selected.

We propose to formulate g as a weighted maximum K-set cover [43]. Let us define a collection of sets $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_N\}$, where $\mathcal{S}_i = \{(w_{kl}, w_{lk}) \,|\, 1 \leq k \leq i < l \leq N\}$ models a



(a) EVM (∞-SC) [10] (b) EVM (10-SC) [10] (c) iEVM (10-wSC) (d) C-EVM (∞-SC) [41] (e) C-iEVM (10-wSC)

Fig. 2. Decision boundaries of different EVM reductions on a 3-class toy dataset. Solid dots correspond to the extreme vectors (EVs) and colored areas belong to the related class where the inclusion probability is visualized via the opacity. In (a), no reduction is performed, *i. e.*, the EVs match the training data. The set cover (SC) reduction and our weighted (wSC) are shown in (b) and (c), respectively. In (d), the C-EVM is shown and (e) presents the C-EVM with our wSC.

```
1: function REDUCE(Θ, K)
2:     Θ_E ← ∅
3:     for k = 1 to K do
4:         idx ← arg max_{i∈|Θ|} Σ_{j=1}^{|Θ|} Ψ_i(x_j)
5:         Θ_E.insert(θ_idx)
6:         Θ.remove(θ_idx)
7:     end for
8:     return Θ_E
9: end function
```

Alg. 1. The proposed weighted maximum K-set cover EVM model reduction.

single EV. A pair $(w_{kl}, w_{lk}) \in \mathcal{S}_i$ contains two weights given by the inclusion probabilities $w_{kl} = \Psi_k(\boldsymbol{x}_l)$ and $w_{lk} = \Psi_l(\boldsymbol{x}_k)$. We determine g according to the integer linear program:

$$\text{maximize} \quad \sum_{i=1}^{N} \sum_{j=i+1}^{N} \beta_{ij} \Psi_i(\boldsymbol{x}_j) + \beta_{ji} \Psi_j(\boldsymbol{x}_i) \tag{6}$$

$$\text{subject to} \quad \sum_{i=1}^{N} \gamma_i \leq K \ , \tag{7}$$

$$\beta_{ij} + \beta_{ji} \leq 1 \ , \tag{8}$$

where $\beta_{ij} \in \{0,1\}$ selects covered elements ($\beta_{ij} = 1 \Leftrightarrow (w_{ij}, w_{ji})$ is covered by $\mathcal{S}_i$) and $\gamma_i \in \{0,1\}$ selects kept EVs ($\gamma_i = 1 \Leftrightarrow \mathcal{S}_i$ is kept). The objective in (6) is optimized w. r. t. β and γ to maximize the value of the coverage. The constraint in (7) limits the amount of EVs to the budget K and (8) penalizes the selections of bilateral coverage.

b) Incremental Algorithm: We solve (6) – (8) by greedy iterations as depicted in Algorithm 1. Our algorithm facilitates incremental learning by reusing intermediate results from the model reduction of the previous epoch, where Θ denotes the intermediate model of a class from the partial fit function and K is the EV budget. Line 3 limits the amount of iterations to the desired budget K. In each iteration, we first compute for each sample the sum of inclusion probabilities from all other samples toward it (line 4). The element with the highest sum is selected as EV (line 5 - 6). In the end, the reduced model Θ_E is released. Note that summations in line 4 do not need to be recomputed in every iteration. We provide additional implementation details for Algorithm 1 in the supplementary material.

c) Relationship to Previous Works [10], [41]: Our weighted maximum K-set cover formulation in (6) – (8) generalizes the conventional set cover model reduction of Rudd *et al.* [10]. To formulate [10] in our framework, we need to substitute $\Psi_i(\boldsymbol{x}_j)$ and $\Psi_j(\boldsymbol{x}_i)$ in (6) by $I(\theta_i)$ and $I(\theta_j)$, *i. e.*, the indicator function of (4). Thus, all samples with coverage probabilities $\geq \zeta$ are weighted uniformly.

The C-EVM [41] uses class-wise DBSCAN clustering [44] and generates centroids from these clusters. This preconditioning reduces the training set size before the actual EVM is fitted to the centroids. However, this does not enforce a specific amount of EVs. This is sub-optimal in memory-limited applications, *e. g.*, on edge devices, where fixed model sizes are preferred.

In Figure 2, we compare different reduction techniques on example data, where K-set cover (K-SC) represents Rudd's method [10] and (K-wSC) our weighted K-set cover (K-wSC). It can be observed that K-SC leads to scattered decision boundaries and is sensitive to outliers. Our stand-alone incremental EVM (iEVM) is robust against outliers and empowers the open space, *cf*. Figure 2c. The C-EVM generates new centroids but does not guarantee a certain amount of EVs. Therefore, we extend it with our K-wSC and bilateral coverage regularization. This selects EVs that accurately describe the underlying distributions of known classes. We argue that both, the iEVM and C-iEVM, perfectly describe different levels of the stability-plasticity tradeoff. While the iEVM strictly bounds the decision boundaries to dense class centers and leaves more open space, it is stable to concept drift. In contrast, the C-iEVM enables more plasticity as outliers have a high impact on the generated centroids.

The hard thresholding of Rudd *et al.* [10] also comes at the cost of embedding their set cover into a bisection search to determine a coverage threshold ζ providing the desired number of EVs. Given a bisection termination tolerance of ϵ, the overall model reduction has a time complexity of $\mathcal{O}(\log(\epsilon^{-1})N^2)$ for a single class comprising N samples. In contrast, our model reduction method avoids thresholding and considers the given budget on the number of EVs in a single pass with time complexity $\mathcal{O}(N^2)$. This is an important factor for implementations on resource limited devices.

V. Open World Evaluation Protocols

We introduce our two designed open world evaluation protocols. The first protocol describes the very general real-world online learning environment, where new classes are

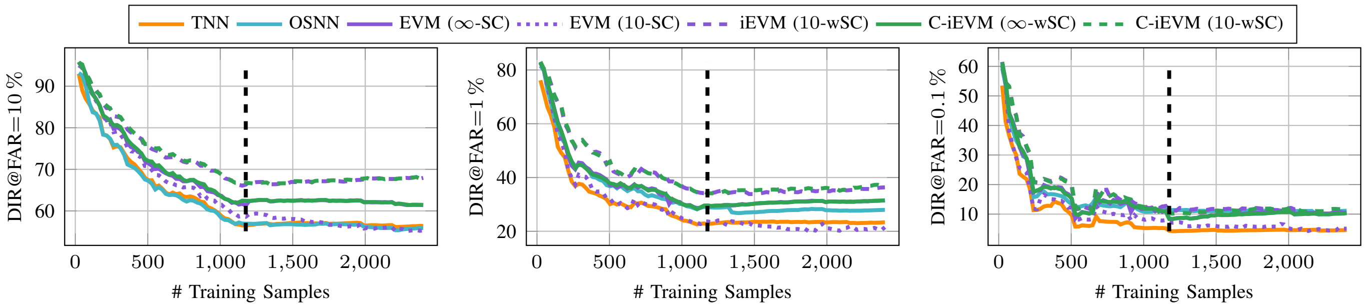


Fig. 3. Averaged results over 3 runs of the proposed open world Protocol I on CIFAR-100. Set cover and our weighted maximum K-set cover reduction to K extreme vectors (EVs) are denoted as K-SC and K-wSC, respectively. The vertical dashed line determines the batch at which the openness remains constant.

learned and old classes are updated by new samples. The second protocol is a specialization of the first one, where subsequent epochs contain only new classes.

1) Protocol I: This protocol reflects the realization of a newly deployed OWR application. While others start with a large initial training phase [1], we argue that this is not possible in real-world scenarios, as the exact environmental conditions, *e. g.*, sensors and lighting, are unknown. Furthermore, it is an unrealistic assumption to start with a large initial training phase.

We start with a minimum of 2 classes and incrementally learn new classes, while incorporating new samples of previous classes. This introduces two types of concept drifts, termed *direct* and *implicit* concept drift. Direct concept drift applies to a single changing class, *e. g.*, the aging of a person. Implicit concept drift determines the mutual impact of neighboring classes competing for transitional feature space. Here, the occurrence of a new class can have a high impact on previously learned classes as both may share parts of the feature space, *e. g.*, leopards and jaguars. Implicit concept drift is given whenever an altering class influences the learned concepts of other classes.

Our protocol allows the control of its complexity on the basis of an initial *openness* [2]. According to this openness, classes are divided into two disjoint sets of knowns $\mathcal{C}_{\mathrm{K}}$ and unknowns $\mathcal{C}_{\mathrm{U}}$, with $|\cdot|$ denoting the cardinality. The first epoch contains 2 classes of $\mathcal{C}_{\mathrm{K}}$. The following epochs comprise a single new class of $\mathcal{C}_{\mathrm{K}}$ as well as samples of classes seen in previous epochs. Hence, all classes in $\mathcal{C}_{\mathrm{K}}$ are known at epoch $|\mathcal{C}_{\mathrm{K}}| - 1$. Each learning epoch follows an evaluation on a fixed test set. Note that, although the test set is fixed, the amount of unknowns reduces over the epochs. Thus, the openness decreases from epoch number 1 to $|\mathcal{C}_{\mathrm{K}}| - 1$. This reduces the complexity of unknown detection while increasing the difficulty for the classification of knowns. To further investigate the models' incremental adaptability at a steady openness, we continue the epoch-wise training after $|\mathcal{C}_{\mathrm{K}}| - 1$ with batches of $\mathcal{C}_{\mathrm{K}}$.

2) Protocol II: This protocol specializes the first one for applications with few samples per class. Due to the limited amount of training samples, we derive a pure class-incremental evaluation, where each epoch contains a certain amount of new classes. No previously learned classes are directly updated by new samples in subsequent epochs but they are updated implicitly by new occurring classes leading to the previously mentioned implicit concept drift.

We split the classes w. r. t. a predefined openness into knowns and unknowns. The unknowns are put in the test set together with a subset of samples for each of the known classes. The known classes are split into batches where each batch contains all remaining samples of a certain amount of classes.

3) Performance Measures: The Detection and Identification Rate (DIR) at certain False Alarm Rates (FARs) serves as evaluation metric, which is common in the open set face recognition [3]. The FAR determines the fraction of misclassified unknowns. The threshold to receive a certain FAR can be derived from the evaluated dataset. The DIR determines the fraction of correctly detected knowns *and* their correct classification. A high DIR at low FAR is favorable.

VI. Experiments and Results

We evaluate our iEVM in different OWR applications. The EVM, OSNN, and Thresholded NN (TNN) serve as baselines. We also extend the C-EVM by our incremental framework, where clustering is applied prior to model fitting. The method notations are adopted from Section IV-3c. Model reductions are performed at every epoch.

1) Image Classification: The open world performance of our approach is evaluated with Protocol I on CIFAR-100 [45]. This dataset comprises 50 000 training and 10 000 test samples of 100 classes. The randomized split into knowns and unknowns is 50 %, which results in an openness range from 80.2 % for the first batch to 18.4 % for batch 49 and the following ones. We evaluate 100 epochs using a batch size of 24 and benchmark all models on the whole test set after each epoch. We repeat the protocol 3 times using different random orders in the creation and processing of batches.

a) Implementation Details: For feature extraction, we use EfficientNet-B6 [46] pre-trained on ImageNet [47] and fine-tuned on a CIFAR-100 training split via categorical cross-entropy loss and a bottleneck layer of size 1024. All EVMs use the same parameters: $\tau = 75$ and $\alpha = 0.5$. For the clustering in the C-EVM and C-iEVM, we adopt the parameters reported in [41]. Methods that employ a model reduction reduce the amount of EVs to $K = 10$. We report additional results with alternative parameters in the supplementary material.

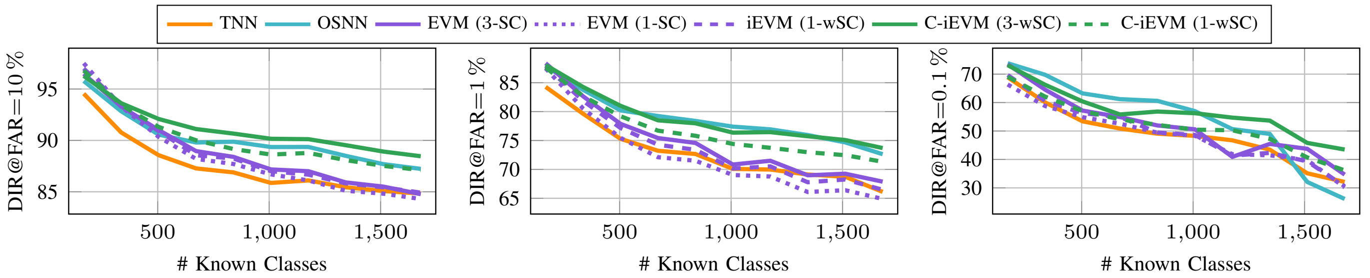


Fig. 4. Averaged results over 3 runs of Protocol II on LFW. Set cover and our weighted maximum K-set cover reduction to K extreme vectors (EVs) are denoted as K-SC and K-wSC, respectively. Our reduction achieves comparable results *while* reducing the model complexity by factor 4.

b) Results: Averaged results of 3 repetitions of Protocol I are shown in Figure 3. We depict the DIR over the amount of samples at different FARs. All EVMs perform similar for the first 250 samples and achieve an initial DIR of about 95 % at a FAR of 10 %. In later epochs, our iEVM and C-iEVM clearly outperform the competing methods for high and medium FARs (10 % and 1 %), while at very small FAR (0.1 %) all methods perform comparably. However, our methods begin to recover after the openness remains constant.

In the case that the training samples within a class are widely spread, the original set cover model reduction struggles to find the most important EVs. This leads to a constant decrease in the DIR even after the openness complexity stays constant. Similarly, DBSCAN in the C-EVM fails to generate meaningful centroids resulting in almost identical outputs as the baseline EVM. We noticed that DBSCAN achieves only average reductions of about 3 % and the model contains 2294 EVs after the last epoch. Our weighted K-set cover easily selects the most important EVs and achieves the best results in the C-iEVM and iEVM while storing only 500 EVs (10 per class).

The amount of EVs does not only influence the memory but also the inference time. The reduced models take about 2.4 s to evaluate the test set while the others require about 14.7 s which is a factor of 6. Further, our model reduction is, averaged over all epochs, by a factor 4.2 faster than the conventional one.

2) Face Recognition: To evaluate our method in open world face recognition, we apply Protocol II to the Labeled Faces in the Wild (LFW) [48], [49] dataset. We adopt the training and the $O3$ test split of [3], where the training set consists of 2900 samples from 1680 unbalanced classes with either 1 or 3 images. We divide this split into 10 batches with 168 classes each. After each epoch the test set is evaluated. Since the test set is highly unbalanced with 1 to 527 samples per class, we report the *macro* average DIR at certain FARs. This prevents the suppression of misclassified underrepresented classes and is therefore a better representation on the global performance on this dataset. The protocol is repeated 3 times.

a) Implementation Details: For feature extraction we use the ResNet50, pre-trained on MS-Celeb-1M [50] and fine-tuned on VGGFace2 [51], with an embedding size of 128. We adopt the EVM parameters $\tau = 75$ and $\alpha = 0.5$ from [3]. Additionally, our methods with model reduction perform the contraction to a single EV per class, *i. e.*, $K = 1$.

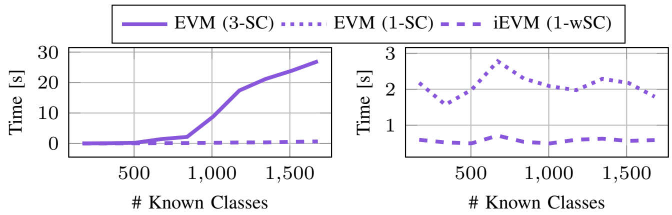


Fig. 5. Averaged runtime of the training step (left) and model reduction (right) from the evaluation of Protocol II and LFW. Our partial fit reduces the average training time by a factor of 28. Our model reduction, averaged over all epochs, is faster than the conventional set cover by a factor of 3.7.

b) Results: We present the averaged DIR at several FARs in Figure 4. Surprisingly, the OSNN achieves in this protocol better recognition scores than in the previous one. The C-EVM and OSNN perform comparable while the OSNN looses precision at the lowest FAR (0.1 %). Our C-iEVM and iEVM achieve comparable results *while* reducing the model complexity by a factor of 4.

The computational efficacy of our incremental framework is presented in Figure 5. Here, partial fitting reduces the average training time by a factor of 28. In particular, performance gains are substantial at late epochs, where the EVM requires 27 s to learn the final classes, while the iEVM takes 0.7 s. Our model reduction is, averaged over all epochs, by a factor of 3.7 faster than the conventional set cover approach.

3) Additional Experiments: The supplementary material contains additional details about the proposed reduction and the evaluation on an additional dataset [52] using Protocol II.

VII. Conclusion

We introduced an incremental leaning framework for the EVM. Our partial model fitting neglects unaffected space during an update and prevents costly Weibull estimates. The proposed weighted maximum K-set cover model reduction guarantees a fixed-size model complexity with less computational effort than the conventional set cover approach. Our reduction leads to dense class centers filtering out outliers. The proposed modifications outperform the original EVM and the C-EVM on novel open world protocols in terms of efficacy and efficiency. In future work, we will investigate the method on larger datasets to better understand the advantages of our model reduction and put more effort into applications with harsh constraints on low False Alarm Rates.

REFERENCES

[1] A. Bendale and T. Boult, "Towards Open World Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1893–1902. 1, 2, 5

[2] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward Open Set Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 35, no. 7, pp. 1757–1772, 2012. 1, 5

[3] M. Günther, S. Cruz, E. M. Rudd, and T. E. Boult, "Toward Open-Set Face Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 71–80. 1, 2, 5, 6, 9, 10

[4] A. Bendale and T. E. Boult, "Towards Open Set Deep Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1563–1572. 1, 2

[5] J. Henrydoss, S. Cruz, E. M. Rudd, M. Günther, and T. E. Boult, "Incremental Open Set Intrusion Recognition Using Extreme Value Machine," in *16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2017, pp. 1089–1093. 1, 2

[6] D. S. Prijatelj, S. Grieggs, F. Yumoto, E. Robertson, and W. J. Scheirer, "Handwriting Recognition with Novelty," in *Document Analysis and Recognition (ICDAR)*, vol. 12824. Springer, 2021, pp. 494–509. 1

[7] B. Lorch, A. Maier, and C. Riess, "Reliable JPEG Forensics via Model Uncertainty," in *IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2020, pp. 1–6. 1, 2

[8] A. Maier, B. Lorch, and C. Riess, "Toward Reliable Models for Authenticating Multimedia Content: Detecting Resampling Artifacts with Bayesian Neural Networks," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 1251–1255. 1

[9] B. Lorch, F. Schirrmacher, A. Maier, and C. Riess, "Reliable Camera Model Identification Using Sparse Gaussian Processes," *IEEE Signal Processing Letters (SPL)*, vol. 28, pp. 912–916, 2021. 1, 2

[10] E. M. Rudd, L. P. Jain, W. J. Scheirer, and T. E. Boult, "The Extreme Value Machine," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 40, no. 3, pp. 762–768, 2017. 1, 2, 3, 4, 9, 10

[11] G. Ditzler, M. D. Muhlbaier, and R. Polikar, "Incremental Learning of New Classes in Unbalanced Datasets: Learn++.UDNC," in *International Workshop on Multiple Classifier Systems (MCS)*. Springer, 2010, pp. 33–42. 1

[12] Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, and Y. Fu, "Large Scale Incremental Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 374–382. 1, 2

[13] G. A. Carpenter and S. Grossberg, "ART 2: Self-Organization of Stable Category Recognition Codes for Analog Input Patterns," *Applied Optics*, vol. 26, no. 23, pp. 4919–4930, 1987. 1

[14] A. Bifet and R. Gavalda, "Adaptive Learning from Evolving Data Streams," in *International Symposium on Intelligent Data Analysis (IDA)*. Springer, 2009, pp. 249–260. 1, 2

[15] G. Cauwenberghs and T. Poggio, "Incremental and Decremental Support Vector Machine Learning," *Advances in Neural Information Processing Systems (NIPS)*, pp. 409–415, 2001. 1, 2

[16] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental Classifier and Representation Learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2001–2010. 1, 2

[17] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari, "End-to-End Incremental Learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 233–248. 1, 2

[18] P. Domingos and G. Hulten, "Mining High-Speed Data Streams," in *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2000, pp. 71–80. 2

[19] R. Polikar, L. Upda, S. S. Upda, and V. Honavar, "Learn++: An Incremental Learning Algorithm for Supervised Neural Networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 31, no. 4, pp. 497–508, 2001. 2

[20] T.-K. Kim, S.-F. Wong, B. Stenger, J. Kittler, and R. Cipolla, "Incremental Linear Discriminant Analysis Using Sufficient Spanning Set Approximations," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007, pp. 1–8. 2

[21] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum Learning," in *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, 2009, pp. 41–48. 2

[22] M. Kumar, B. Packer, and D. Koller, "Self-Paced Learning for Latent Variable Models," in *Advances in Neural Information Processing Systems (NIPS)*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., vol. 23. Curran Associates, Inc., 2010. 2

[23] L. Lin, K. Wang, D. Meng, W. Zuo, and L. Zhang, "Active Self-Paced Learning for Cost-Effective and Progressive Face Identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 40, no. 1, pp. 7–19, 2017. 2

[24] C. Zhang, N. Song, G. Lin, Y. Zheng, P. Pan, and Y. Xu, "Few-Shot Incremental Learning with Continually Evolved Classifiers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 455–12 464. 2

[25] D. M. Tax and R. P. Duin, "Growing a Multi-Class Classifier with a Reject Option," *Pattern Recognition Letters*, vol. 29, no. 10, pp. 1565–1570, 2008. 2

[26] P. L. Bartlett and M. H. Wegkamp, "Classification with a Reject Option Using a Hinge Loss," *Journal of Machine Learning Research (JMLR)*, vol. 9, no. 8, pp. 1823–1840, 2008. 2

[27] Y. Grandvalet, A. Rakotomamonjy, J. Keshet, and S. Canu, "Support Vector Machines with a Reject Option," in *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS)*, 2008. 2

[28] H. Cevikalp and B. Triggs, "Efficient Object Detection Using Cascades of Nearest Convex Model Classifiers," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 3138–3145. 2

[29] W. J. Scheirer, L. P. Jain, and T. E. Boult, "Probability Models for Open Set Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 36, no. 11, pp. 2317–2324, 2014. 2

[30] L. P. Jain, W. J. Scheirer, and T. E. Boult, "Multi-Class Open Set Recognition Using Probability of Inclusion," in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 393–409. 2

[31] S. D. Zongyuan Ge and R. Garnavi, "Generative OpenMax for Multi-Class Open Set Classification," in *Proceedings of the British Machine Vision Conference (BMVC)*, no. 42. BMVA Press, 2017, pp. 1–12. 2

[32] L. Neal, M. Olson, X. Fern, W.-K. Wong, and F. Li, "Open Set Learning with Counterfactual Images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 613–628. 2

[33] S. Kong and D. Ramanan, "OpenGAN: Open-Set Recognition via Open Data Generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 813–822. 2

[34] Z. Yue, T. Wang, Q. Sun, X.-S. Hua, and H. Zhang, "Counterfactual Zero-Shot and Open-Set Visual Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 404–15 414. 2

[35] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight Uncertainty in Neural Network," in *International Conference on Machine Learning (ICML)*. PMLR, 2015, pp. 1613–1622. 2

[36] Y. Sun, C. Guo, and Y. Li, "ReAct: Out-of-Distribution Detection with Rectified Activations," *Advances in Neural Information Processing Systems (NIPS)*, vol. 34, 2021. 2

[37] P. R. M. Júnior, R. M. De Souza, R. d. O. Werneck, B. V. Stein, D. V. Pazinato, W. R. de Almeida, O. A. Penatti, R. d. S. Torres, and A. Rocha, "Nearest Neighbors Distance Ratio Open-Set Classifier," *Springer Machine Learning*, vol. 106, no. 3, pp. 359–386, 2017. 2

[38] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka, "Distance-Based Image Classification: Generalizing to New Classes at Near-Zero Cost," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 35, no. 11, pp. 2624–2637, 2013. 2

[39] M. Ristin, M. Guillaumin, J. Gall, and L. Van Gool, "Incremental Learning of NCM Forests for Large-Scale Image Classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3654–3661. 2

[40] K. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards Open World Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5830–5840. 2

[41] J. Henrydoss, S. Cruz, C. Li, M. Günther, and T. E. Boult, "Enhancing Open-Set Recognition Using Clustering-Based Extreme Value Machine (C-EVM)," in *International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 441–448. 2, 4, 5

[42] R. M. Karp, "Reducibility Among Combinatorial Problems," in *Complexity of Computer Computations*. Springer, 1972, pp. 85–103. 2

[43] R. Cohen and L. Katzir, "The Generalized Maximum Coverage Problem," *Information Processing Letters*, vol. 108, no. 1, pp. 15–22, 2008. 3
[44] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, vol. 96, no. 34. AAAI Press, 1996, pp. 226–231. 4
[45] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," University of Toronto, Tech. Rep., 2009. 5
[46] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 6105–6114. 5
[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255. 5
[48] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007. 6
[49] G. B. Huang and E. Learned-Miller, "Labeled Faces in the Wild: Updates and New Reporting Procedures," University of Massachusetts, Amherst, Tech. Rep. UM-CS-2014-003, May 2014. 6
[50] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 87–102. 6
[51] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A Dataset for Recognising Faces Across Pose and Age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2018, pp. 67–74. 6
[52] S. Fiel, F. Kleber, M. Diem, V. Christlein, G. Louloudis, S. Nikos, and B. Gatos, "ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI)," in *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 1377–1382. 6, 9, 10
[53] V. Christlein, M. Gropp, S. Fiel, and A. Maier, "Unsupervised Feature Learning for Writer Identification and Writer Retrieval," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, Nov 2017, pp. 991–997. 10

APPENDIX

A. Algorithm Details

Algorithm 2 provides additional details of the proposed weighted maximum K-set cover model reduction for the EVM. Recall that this is a class-wise reduction technique. Thus, the amount of EVs in a single class is denoted as E. The amount of samples within a batch of this class is denoted N.

The summations of the inclusion probabilities for each EV are given in $\boldsymbol{p}$. The EVM model Θ_E^t represents the EVs of the previous epoch, Θ_N^{t+1} the estimated Weibull parameters of the current data batch, and K determines the EV budget. The reduction comprises four steps:

1) Updating the inclusion probability sums of the old EVs w. r. t. the new batch (line 2 to 4).
2) Sum up the inclusion probabilities of the new samples w. r. t. each other (line 6 to 9). This step has a time complexity of $\mathcal{O}(N \cdot (E+N))$ which is $\mathcal{O}(N^2)$ for large batches (*i. e.*, $N \gg E$) and $\mathcal{O}(NE)$, otherwise.
3) In line 10 follows the greedy search for the EVs. Details for Algorithm 3 follow in the next paragraph.
4) Update $\boldsymbol{p}$ according to the new EVs (line 11 to 15). If the two conditions $N > E$ and $E > (N - E)$ hold, it is more efficient to skip line 11, *i. e.*, not to reset $\boldsymbol{p}$. Then we can use the modified $\boldsymbol{p}$ of Algorithm 3 and incrementally subtract and remove non-EV samples similar as in the regularization in Algorithm 3. This has a time complexity of $\mathcal{O}((N - E) \cdot E) \Rightarrow \mathcal{O}(NE)$, since we only need to update the elements in $\boldsymbol{p}$ that are part of Θ_E^{t+1}.

The greedy iteration algorithm is depicted in Algorithm 3 and requires the summations $\boldsymbol{p}$, the combined model Θ, and the budget K. The amount of iterations is limited by K (line 3). In line 4 we take the sample with the highest sum of inclusion probabilities and store it in the EV model (line 5). Then follows the bilateral coverage regularization by removing the probability of inclusion of the selected EV from the other samples (line 6 to 8). In line 9 to 10, we remove the EV from $\boldsymbol{p}$ and Θ. In the end, we receive the EVM model Θ_E containing only the EVs. Note for the mentioned special case in the previous step 4, we also need to return the modified $\boldsymbol{p}$ and Θ.

The total asymptotic runtime of the proposed weighted maximum K-set cover algorithm is $\mathcal{O}(N^2)$. It does not depend on a bisection search as the set cover of Rudd *et al.* [10] that has a complexity of $\mathcal{O}(\log(\epsilon^{-1})N^2)$, with termination tolerance ϵ.

1: **function** REDUCE($\boldsymbol{p}^t$, Θ_E^t, Θ_N^{t+1}, K)
   ▷ Update EV sums: $\mathcal{O}(EN)$
2: **for** e **in** E **do**
3: $\boldsymbol{p}[e] \leftarrow \boldsymbol{p}[e] + \sum_{n \in N} \Psi_e(\boldsymbol{x}_n)$
4: **end for**
   ▷ Compute sums for new samples: $\mathcal{O}(N^2)$
5: $\Theta \leftarrow \Theta_E \cap \Theta_N$
6: **for** n **in** N **do**
7: $p \leftarrow \sum_{i \in |\Theta|} \Psi_n(\boldsymbol{x}_i)$
8: $\boldsymbol{p}$.insert(p)
9: **end for**
10: $\Theta_E^{t+1} \leftarrow$ Greedy($\boldsymbol{p}, \Theta, K$)
   ▷ Compute new $\boldsymbol{p}$: $\mathcal{O}(E^2)$
11: $\boldsymbol{p}^{t+1} \leftarrow \emptyset$
12: **for** i **in** $|\Theta_E^{t+1}|$ **do**
13: $p \leftarrow \sum_{j \in |\Theta_E^{t+1}|} \Psi_i(\boldsymbol{x}_j)$
14: $\boldsymbol{p}^{t+1}$.insert(p)
15: **end for**
16: **return** Θ_E^{t+1}, $\boldsymbol{p}^{t+1}$
17: **end function**

Alg. 2. Detailed version of the proposed class-wise weighted maximum K-set cover model reduction for the Extreme Value Machine (EVM).

1: **function** GREEDY($\boldsymbol{p}$, Θ, K)
2: $\Theta_E \leftarrow \emptyset$
3: **for** $k = 1$ **to** K **do** ▷ $\mathcal{O}(KN)$
4: idx $\leftarrow \arg\max \boldsymbol{p}$
5: Θ_E^{t+1}.insert(θ_{idx})
   ▷ Bilateral coverage regularization: $\mathcal{O}(N)$
6: **for** $i = 1$ **to** $|\Theta|$ **do**
7: $\boldsymbol{p}[i] \leftarrow \boldsymbol{p}[i] - \Psi_i(\boldsymbol{x}_{\text{idx}})$
8: **end for**
9: $\boldsymbol{p}$.remove(p_{idx})
10: Θ.remove(θ_{idx})
11: **end for**
12: **return** Θ_E
13: **end function**

Alg. 3. Greedy iterations with bilateral coverage regularization to solve the weighted maximum K-set cover model reduction.

B. Additional Experiments

In this section we present further experiments of the evaluation with Protocol I and CIFAR-100. Furthermore, we evaluated the writer identification dataset ICDAR17 [52] with Protocol II.

1) Protocol I – CIFAR-100: In the main text, we show the result of the iEVM on Protocol I and CIFAR-100 with parameters $\tau = 75$ and the reduction to $K = 10$. Here, we want to present further parameterizations in Figure 6. As in the main text, the left, middle, and right plots show the DIR at FARs of 10 %, 1 %, and 0.1 %.

When comparing the accuracies for different values of τ at identical K, it turns out that the tail size τ has almost no influence on the models' accuracy. This is similar to what Günther *et al.* [3] reported on the LFW dataset.

A larger value of K may lead to worse results, as can be seen in the case of iEVM ($\tau = 75$, 50-wSC). This may be counter-intuitive at first glance, considering that classification should perform better with more data. However, storing more data implies less plasticity and more stability which can interfere with the incremental training adaptability.

2) Protocol II – ICDAR17: Another OWR task is writer identification. Here, we apply Protocol II to the dataset

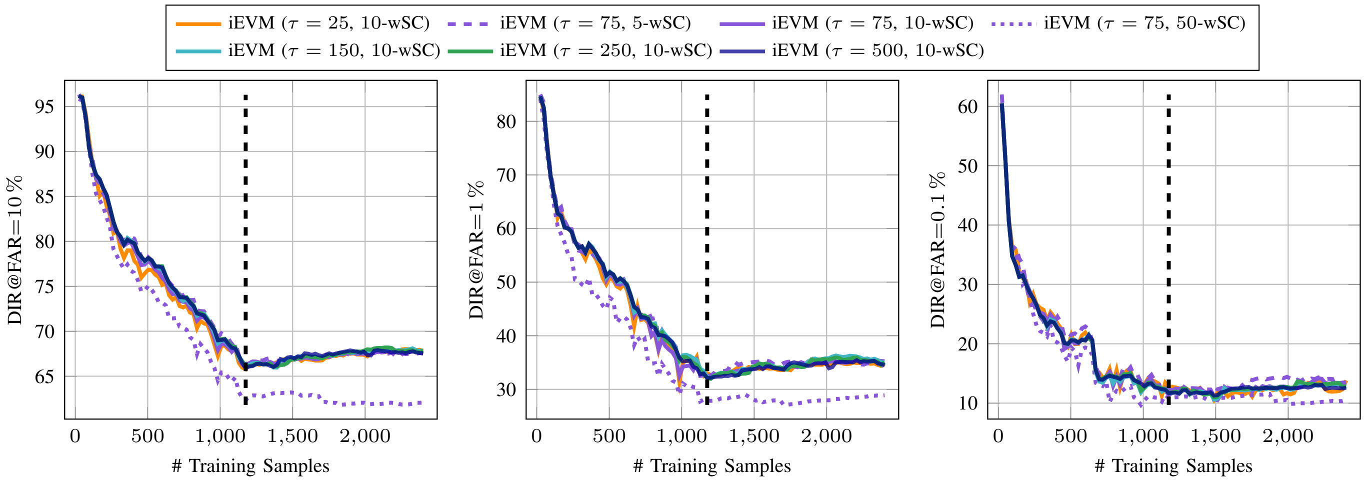


Fig. 6. Different parameterizations of our incremental EVM (iEVM). Averaged results over 3 runs of Protocol I and CIFAR-100. The vertical dashed line determines the batch at which the openness remains constant.

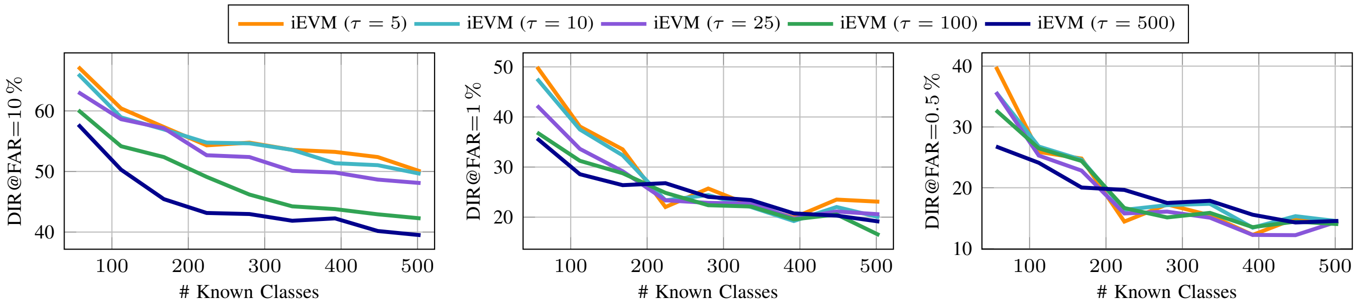


Fig. 7. Different tail size τ parameterizations of our incremental EVM (iEVM). Averaged results over 3 runs of Protocol II and ICDAR17.

ICDAR17 [52]. It contains handwritten pages from the 13th to 20th century. Since the feature extraction is trained on the training set of ICDAR17, the subsequent classification training and evaluation on the same set would be biased. Therefore, we take only the test set into account with 5 pages for each of the 720 writers. 30 % of the classes are selected as unknowns and left in the test split. For each of the known classes, we leave 1 sample in the test split, *i. e.*, the training split has 4 samples for each of the 504 known classes. The knowns are split into 9 batches with 56 classes and trained incrementally. This protocol implements an openness from 62 % to 9.3 %. The results are averaged over 3 protocol repetitions.

a) Implementation Details: The feature set consists of the 6400-dimensional activation of the penultimate layer of a ResNet20. It was trained in a self-supervised fashion [53]. The training uses SIFT descriptors that are calculated on patches of 32×32 pixels at SIFT keypoints. The SIFT descriptors are clustered using k-means. Then, the ResNet20 is trained using cross-entropy loss where the patches are used as input and the targets are the cluster center IDs of the patches.

b) Hyperparameter Evaluation: The experiments on CIFAR-100 and Protocol I show, similar as the previous work of Günther *et al.* [3], that the tail size parameter τ has only a minor impact on the results. However, we noticed that this does not apply to Protocol II and ICDAR17 as visualized in Figure 7. The experiments show that a small tail size ($\tau \in \{5, 10\}$) achieves a better DIR at a high FAR of 10 %. This difference degrades over the class-wise increments at medium and small FARs of 1 % and 0.5 %. Rudd *et al.* [10] state that a larger tail size leads to higher coverage. This implies that for ICDAR17 a high coverage and little open space is less favorable and a steep decision boundary is beneficial.

c) Results: The comparison to the other baseline methods follows in Figure 8. All EVMs use a tail size $\tau = 5$. The C-iEVM without model reduction performs comparable to the OSNN and both outperform the conventional EVM. The boundary case of a model reduction to a single EV per class does not lead to an improvement in this evaluation. In contrast to this result, we note that the evaluation of Protocol I on CIFAR-100 performed much better with model reduction. However, the representation of a class via a single sample is challenging and heavily depends on the class distribution.

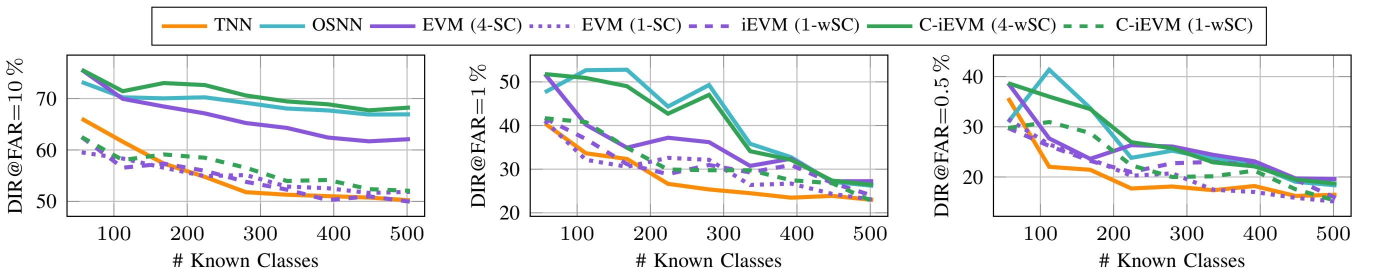


Fig. 8. Averaged results over 3 runs of Protocol II on ICDAR17. Set cover and our weighted maximum K-set cover reduction to K extreme vectors (EVs) are denoted as K-SC and K-wSC, respectively.