Simple Weak Coresets for Non-decomposable Classification Measures

Jayesh Malaviya[1], **Anirban Dasgupta**[1], **Rachit Chhaya**[2]

[1]Indian Institute of Technology, Gandhinagar
[2]DA-IICT, Gandhinagar
malaviya_jayesh@iitgn.ac.in, anirbandg@iitgn.ac.in, rachit_chhaya@daiict.ac.in

Abstract

While coresets have been growing in terms of their application, barring few exceptions, they have mostly been limited to unsupervised settings. We consider supervised classification problems, and non-decomposable evaluation measures in such settings. We show that stratified uniform sampling based coresets have excellent empirical performance that are backed by theoretical guarantees too. We focus on the F1 score and Matthews Correlation Coefficient, two widely used non-decomposable objective functions that are nontrivial to optimize for, and show that uniform coresets attain a lower bound for coreset size, and have good empirical performance, comparable with "smarter" coreset construction strategies.

Introduction

In typical classification tasks, multiple objective functions are often at stake— one is the (surrogate) classification objective being optimized, and the other is the real measure of performance. This "real measure" is non-differentiable and, more importantly, often a non-decomposable function over the set of input points, i.e., it cannot be written as a sum over the input points. Popular examples include F_1 score, Matthews Correlation Coefficient (MCC), area under the curve (AUC-ROC) measures, and various variants of combinations of precision and recall measures e.g. H-mean, G-mean etc. In spite of being non-differentiable and non-decomposable, such classification measures are attractive since they help portray the tradeoffs between precision and recall, help enable the handling of mild to severe label imbalance, etc.

Such non-additive measures are generally optimized empirically by choosing a surrogate decomposable measure, then optimizing the decision boundary threshold using grid search. In recent years, a host of algorithms have been developed for optimizing the non-decomposable measures directly (Joachims 2005; Nan et al. 2012; Narasimhan, Kar, and Jain 2015; Eban et al. 2017). However, the main drawback remains that the algorithms developed are often not very efficient in practice and, thereby, remain harder to scale over large datasets.

For decomposable loss functions, such as for regression or matrix factorization, one way in which this complexity of training has been circumvented is by constructing coresets—summaries of the dataset that enable an optimizer to optimize over the coreset only and give both a theoretically guaranteed as well as empirically satisfying performance for the full data. The major tool for coreset creation remains importance sampling, and that relies crucially on the decomposability of the loss function in order to deduce the importance weights. Indeed, in traditional classification coreset literature, researchers have explored various coreset building mechanisms for logistic loss (Munteanu et al. 2018), SVM (Tukan et al. 2021), etc. However, again the main catch is their evaluation and bounds focus on the decomposable surrogates such as the logistic loss function, hinge loss function, etc. (Mai, Musco, and Rao 2021). For non-decomposable performance measures e.g. F1 score, MCC score, AUC etc. there does not exist any such coreset creation mechanisms.

Our paper is the first step along this direction. In this work, we show that for F1-score and MCC, applying stratified uniform sampling is enough to obtain a (weak) coreset that preserves the value of these measures up to a small additive error for an interesting set of queries that contains the optimal. We also show a lower bound for strong coresets for the F1 and the MCC scores, implying that we cannot do much better than uniform sampling. We provide experimental evidence of our results on real datasets and for various classifiers and sampling techniques. Even though we formally prove our bounds and theory for F1-score and MCC, it applies to most contingency table-based measures.

Following are our main contributions:

1. We provide lower bounds against construction of strong coresets for both MCC and F1 score measures.
2. We show that sampling uniformly from each class in a stratified manner gives a weak coreset for the F1 score and a weak coreset with a small additive error for MCC. Here weak signifies that our coreset works for a set of a large number of 'important' classifiers, including the optimal.
3. We provide empirical results for a number of different classifiers and real data sets comparing uniform sampling with other well-known sophisticated coreset construction strategies.

The rest of the paper is organized as follows: Sections 2 and 3 give the necessary background and related work. Section

4 provides lower bounds(negative results) for coresets for both MCC and F1 scores. Sections 5 and 6 give the analysis of our weak coreset construction guarantees for F1 score and MCC. We discuss the experiments and comparison of uniform sampling on real data sets with a variety of sampling algorithms in section 7 and conclude in section 8. Some proofs and additional experiments are available in the arXiv version [1] .

Background

A coreset can be considered to be a weighted subset of data from the original dataset or some different small-sized representation of the original dataset. In designing a coreset, the main challenge is to build one with provable tradeoffs between the approximation error and the coreset size.

We interchangeably refer to classifiers as "query" and denote them by q. Q denotes the set of all possible classifiers.

Definition 1. *(Coreset) (Agarwal et al. 2005) Given a weighted dataset X, let $x \in X$ and $\mu_X(x)$ be its corresponding nonnegative weight. Let Q be a set of solutions known as query space and $q \in Q$ be a query. For each $q \in Q$, let $f_q : X \to \mathbb{R}_{\geq 0}$ be a non-negative function. Define $cost(X, q) = \sum_{x \in X} \mu_X(x) f_q(x)$. For $\epsilon > 0$, a weighted-set (C, w) is an ϵ-coreset of X for the cost function $\{f_q\}$, if $\forall q \in Q$,*

$$|cost(X, q) - cost(C, q)| \leq \epsilon \; cost(X, q).$$

For coresets with a θ additive error, the above guarantee becomes the following– for all q.

$$|cost(X, q) - cost(C, q)| \leq \epsilon \; cost(X, q) + \theta.$$

Though stated in terms of decomposable loss functions, this definition also applies to measures which are not decomposable but are rather functions of entire dataset as a whole. We will also be dealing with weak coresets, i.e. our coresets will satisfy the above guarantees for a specific subset of the solution space, that contains the optimal.

In this text, we mainly deal with uniform sampling, and with two non-decomposable loss functions, the F1 score and the MCC score. For a particular classifier q, let $tp(q), tn(q), fp(q)$ and $fn(q)$ denote the true-positive, the true-negative, the false positive and false negative respectively. We will sometimes drop the (q) notation when it is clear from context.

The F1 score is defined as the following. For a classifier q,

$$F_1(q) = 2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{tp}{tp + \frac{1}{2}(fp + fn)}$$

The MCC score is defined as the following (we leave off the "(q)" for lack of space).

$$MCC(q) = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} = \frac{\frac{tp}{n} - T'P'}{\sqrt{T'P'(1 - T')(1 - P')}}$$

[1] http://arxiv.org/abs/2312.09885

where T' denotes the fraction of ground truth positives and $P' = (tp + fp)/n$ denotes the fraction of predicted positives. We will use Y^+ and Y^- to denote the set of all positive and negatively labeled points. Sometimes we will also use the same notation to denote the total number of positive and negatively labeled points.

Related Work

(Joachims 2005) formulated the non-decomposable measures as a structural prediction problem and gave a direct optimization algorithm using multivariate SVM formulation. The primary idea is to reduce the number of constraints in multivariate SVM by finding the most violated constraint in each iteration, preparing the candidate set, and applying multivariate SVM over that small set only. But in their algorithm for calculation of the argmax for most violated constraints, the algorithm must go over all possible configurations of the contingency table, which amount to n^2 different contingency tables.

(Kar, Narasimhan, and Jain 2014) extend the existing online learning models for point-loss functions to non-decomposable loss functions. They also develop scalable SGD solvers for non-decomposable loss functions. Their work uses the previous result of representing the non-decomposable loss function as structural SVMs (Joachims 2005) and optimizing them. Deep learning based methods for optimizing non-decomposable losses have also been explored in (Sanyal et al. 2018). We note that our work on coresets is orthogonal to the work on optimizing such losses, any such optimizer could be used in parallel with our coreset technique.

(Eban et al. 2017) proposed an alternative formulation based on simple direct bounds on per-sample quantities indicating whether each sample is a true positive or a false positive. Using these bounds, they constructed global bounds on ranking-based measures such as precision at fixed recall, recall at fixed precision, and F1 score, among others. From these global bounds, they derived the surrogate objective function and the closed-form alternate optimization problem, which is then optimized. However, they did not discuss empirical results for the F1 score and MCC. (Bénédict et al. 2022) try to maximize a surrogate loss in place of the F1 score. There has been some work in active sampling to estimate F1 score using optimal subsampling (Sawade, Landwehr, and Scheffer 2010) or iterative importance sampling (Poms et al. 2021) ,however, both the motivation and guarantees are different from the coreset guarantees.

Coresets have been studied since (Agarwal et al. 2005). One of the popular ways to create coreset is to construct a probability distribution, called *sensitivity*, over the set of input points (Langberg and Schulman 2010; Feldman and Langberg 2011). Sampling points proportional to their sensitivities and appropriately reweighing them gives a coreset with high probability (Langberg and Schulman 2010; Feldman and Langberg 2011). However, the major technicalchallenge in this method is to figure out computationally inexpensive upper bounds to the sensitivity scores for each cost function. More details about the construction of coresets using the sensitivity framework can be found in (Feldman

and Langberg 2011; Bachem, Lucic, and Krause 2017, 2018; Braverman et al. 2016; Mai, Musco, and Rao 2021; Feldman 2020) and references within. While coresets using this framework have theoretical guarantees, it has been shown that for some problems you can not do much better than uniform sampling (Samadian et al. 2020). Our paper also has similar results. It has been shown by (Braverman et al. 2022) that using uniform sampling with some careful analysis can also give coreset guarantees for clustering problems. Additionally empirically uniform sampling has been shown to be comparable to other techniques for many real data sets (Lu, Raff, and Holt 2023).

Lower Bounds

In this section we first present lower bounds for strong coresets for F_1 score and MCC.

Lower Bound for Strong F_1 Coresets

Theorem 1. *Let $F_1(q)$ be the F_1 score on full dataset D with fixed query q and $\tilde{F_1}(q)$ be the F_1 score on coreset C. Then, there does not exist a strong coreset C with a size less than n that satisfies relative error approximation i.e. $\tilde{F_1}(q) \in (1 \pm \epsilon)F_1(q)$ for all queries $q \in Q$.*

Proof. Let d be any non-negative even integer. Let $\mathcal{S}$ be the set of all sets of size $d/2$. We define the set P of points as follows: In order to generate each point p, choose $B_p \in \mathcal{S}$, and $y_p \in \{\pm 1\}$. The vector p is defined as following— $p_i = -y_p$ for all $i \in B_p$, and $p_i = 0$ for all $i \in [d] \setminus B_p$, $p_{d+1} = y_p/2$. The label of the point p is also y_p.

For each point p, we also define a corresponding classifier w_p in the following way— $w_p[i] = 0$ if $i \in B_p$, $w_p[i] = 1$ if $i \in [d] \setminus B_p$, and $w_p[d+1] = 1$.

By the above construction, for every point

$$p \cdot w_p = y_p/2.$$

Hence $sign(p \cdot w_p) = y_p$, and hence p is correctly classified. For any point $q \neq p$, $B_p \neq B_q$. Hence,

$$q \cdot w_p = -y_q |B_q \cap \bar{B_p}| + y_q/2$$

Since $|B_q \cap \bar{B_p}| \geq 1$, $sign(q \cdot w_p) = -y_q$, i.e. q is misclassified.

So for a point p with $y_p = +1$, for the query w_p the F_1 score is $\frac{1}{1+(n-2)/2}$. However, if the coreset does not sample the point p, the F_1 score for w_p would be zero, which is not a relative error approximation. Since this is true for every point p, the coreset has to be of size n.

□

Lower Bound for Strong MCC Coresets

Theorem 2. *Let $MCC(q)$ be the MCC score on full dataset D with fixed query q and $\widetilde{MCC}(q)$ be the MCC score on coreset C. Then, there does not exist a strong coreset C with a size less than n that satisfies relative error approximation i.e. $\widetilde{MCC}(q) \in (1 \pm \epsilon)MCC(q)$ for all queries $q \in Q$.*

Proof. Let $d \geq 4$ be any non-negative even integer. Let $\mathcal{S}$ be the set of all sets of size $d/2$. Let $n = \binom{d}{d/2}$. Let $y_p \in \{\pm 1\}$ be such that exactly $n/2$ points have $y_p = +1$. We define the set P of n points as follows: in order to generate each point $p \in \mathbb{R}^{d+1}$, choose $B_p \in \mathcal{S}$. The vector p is defined as follows: $p_i = X_p$ for all $i \in B_p$, and $p_i = 0$ for all $i \in [d] \setminus B_p$, $p_{d+1} = y_p/2$. The label of the point p is also y_p.

Now, lets define X_p for points, $1 \leq p \leq n$ as follows. If the true label of point p is $y_p \in Y^+$ then, $X_p = y_p$ for $1 \leq p \leq \frac{Y^+}{2}$, and $-y_p$ else. Similarly, for $p \in Y^-$, $X_p = y_p$ for $1 \leq p \leq \frac{Y^-}{2}$ and $-y_p$ else.

For each point p, we also define a corresponding classifier w_p in the following way— $w_p[i] = 0$ if $i \in B_p$, $w_p[i] = 1$ if $i \in [d] \setminus B_p$, and $w_p[d+1] = 1$.

By the above construction, for every point

$$p \cdot w_p = y_p/2.$$

Hence $sign(p \cdot w_p) = y_p$, and hence p is correctly classified. For any point $q \neq p$, $B_p \neq B_q$. Hence,

$$q \cdot w_p = X_q |B_q \cap \bar{B_p}| + y_q/2$$

Since $|B_q \cap \bar{B_p}| \geq 1$, $sign(q \cdot w_p) = X_q$, i.e. q is classified as per the sign of X_q.

Therefore, for the classifier w_p, all of the points, except point p, classify according to X_q.

Notice that we have designed X_p for all p such that for Y^+ and Y^-, it will classify half points positive and half points as negative.

Therefore for a point p with $y_p = +1$ and classifier w_p, we have $tp = \frac{n-1}{4} + 1$, $tn = \frac{n-1}{4}$, $fn = \frac{n-1}{4}$ and $fp = \frac{n-1}{4}$ respectively for balanced setup i.e. $Y^+ = Y^- = \frac{n}{2}$.

For above setup MCC score is,

$$\begin{aligned} MCC &= \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}} \\ &= \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp+fp)Y^+Y^-(tn+fn)}} \\ &= \frac{(\frac{n-1}{4}+1)\cdot(\frac{n-1}{4}) - (\frac{n-1}{4})\cdot(\frac{n-1}{4})}{\frac{n}{2}\sqrt{(\frac{n+1}{2})(\frac{n-1}{2})}} \\ &= \theta\left(\frac{1}{n}\right) \end{aligned}$$

However, if the coreset does not sample the point p, then we have $tp = \frac{n-1}{4}$, $tn = \frac{n-1}{4}$, $fn = \frac{n-1}{4}$ and $fp = \frac{n-1}{4}$ respectively, thus MCC score for w_p would be zero, which is not a relative error approximation. Since this is true for every point p, the strong coreset has to be of size n.

□

Weak Coreset for Queries with High F_1 Score

We will be using the following result of (Li, Long, and Srinivasan 2001) to get our coreset guarantees for all queries in the query set of our interest.

Theorem 3. *(Li, Long, and Srinivasan 2001)*

Let, $\alpha > 0$, $v > 0$ and $\delta > 0$. Fix a countably infinite domain X and let $q(\cdot)$ be any probability distribution over X. Let F be a set of functions from X to $[0,1]$ with $Pdim(F) = d^{'}$. Denote by C a sample of m points from X sampled independently according to $q(\cdot)$. Then, for $m \in \Omega\left(\frac{1}{\alpha^2 \cdot v}\left(d^{'}\log(1/v) + \log(1/\delta)\right)\right)$, with probability at least $1-\delta$ it holds that,

$$\forall f \in F; d_v\left(\sum_{x\in X} q(x)f(x), \frac{1}{|C|}\sum_{x\in C} f(x)\right) \leq \alpha$$

where, $d_v(a,b) = \frac{|a-b|}{a+b+v}$.

Now we state our theorem for bounding the coreset size and performance for F_1 score.

Theorem 4. *Let $\epsilon > 0$ and $c > 1$. Consider an instance where number of positive samples are Y^+ and number of negative samples are Y^-, and $n = Y^+ + Y^-$. We consider Q_γ to be the set of queries such that $F_1(q) \geq \gamma$ and $tp \geq max\left(\frac{n(1-c\cdot\epsilon)}{2c\cdot(1-\epsilon)}, \frac{n(1+c\cdot\epsilon)}{2c\cdot(1+\epsilon)}\right)$ for $q \in Q_\gamma$. Let $d = vc\text{-}dimension(Q_\gamma)$.*

Stratified uniform sampling with a total of $\left(\frac{(2-\gamma)^2}{\gamma^2\epsilon^2} + \frac{1}{\epsilon^2} + \left(\frac{Y^-}{Y^+}\right)^2\frac{1}{\epsilon^2}\right) \cdot \left(d + \log\frac{1}{\delta}\right)$ samples would be able to give a coreset for Q_γ that satisfies $\widetilde{F_1}(q) \in (1 \pm c\cdot\epsilon)F_1(q)$ for all queries $q \in Q_\gamma$ with probability at least $1-3\delta$, for a suitable $c > 1$.

Proof. We have, $\frac{tp}{tp+\frac{1}{2}(fp+fn)} = F_1 \geq \gamma$, and hence $\frac{tp}{\frac{1}{2}(fp+fn)} \geq \frac{\gamma}{1-\gamma}$. Also, $tp + fp + fn \geq Y^+$. We now find the minimum value of tp satisfying these two inequalities.

From above, $tp = \left(\frac{\gamma}{1-\gamma}\right)\cdot\frac{1}{2}(fp+fn) + k$, where $k \geq 0$. Then, $\left(\frac{\gamma}{1-\gamma}\right)\frac{1}{2}(fp+fn) + k + (fp+fn) \geq Y^+$ and hence $\frac{1}{2}(fp+fn) \geq (Y^+ - k)/\left(\frac{\gamma}{1-\gamma}+2\right)$. Thus $tp = \frac{\left(\frac{\gamma}{1-\gamma}\right)Y^+}{\left(\frac{\gamma}{1-\gamma}+2\right)} + \frac{2k}{\frac{\gamma}{1-\gamma}+2} \geq \frac{\gamma\cdot Y^+}{(2-\gamma)}$.

Let $\widetilde{tp}$ and $\widetilde{fn}$ be the true positive and false negative counts obtained from the sampled set. Let all the negative samples be given a weight of $\frac{Y^-}{Y^+}$. Hence, let $\widetilde{\widetilde{fp}}$ defined as following—$\widetilde{\widetilde{fp}} = \left(\frac{Y^-}{Y^+}\right)\widetilde{fp}$. Similarly, let $\widetilde{\widetilde{tn}}$ defined as following—$\widetilde{\widetilde{tn}} = \left(\frac{Y^-}{Y^+}\right)\widetilde{tn}$. Our algorithm uses the four estimates $\widetilde{tp}$, $\widetilde{\widetilde{fp}}$, $\widetilde{\widetilde{tn}}$ and $\widetilde{fn}$ to estimate the F1 score. We first show the quality of each of these estimates by applying Theorem 3 individually.

For tp approximation,

$$\forall q \in Q; d_v\left(\sum_{x\in Y^+}\frac{1}{Y^+}\delta_x, \frac{1}{|S_1|}\sum_{x\in S_1}\delta_x\right) \leq \alpha_1$$

where, $\delta_x = 1(x \in tp)$ is indicator variable and $d_v(a,b) = \frac{|a-b|}{a+b+v}$

$$= \frac{\left|\sum_{x\in Y^+}\frac{1}{Y^+}\delta_x - \frac{1}{|S_1|}\sum_{x\in S_1}\delta_x\right|}{\sum_{x\in Y^+}\frac{1}{Y^+}\delta_x + \frac{1}{|S_1|}\sum_{x\in S_1}\delta_x + v} \leq \alpha_1$$

$$= \frac{\left|\frac{tp}{Y^+} - \frac{\widetilde{tp}}{|S_1|}\right|}{\frac{tp}{Y^+} + \frac{\widetilde{tp}}{|S_1|} + v} \leq \alpha_1$$

$$= \left|\frac{tp}{Y^+} - \frac{\widetilde{tp}}{|S_1|}\right| \leq 3\alpha_1$$

Since, $\frac{tp}{Y^+}$ and $\frac{\widetilde{tp}}{|S_1|}$ is less than one, lets take $v = \frac{1}{2}$.

$$\widetilde{tp} \in tp\left(\frac{S_1}{Y^+}\right) \pm S_1\cdot(3\alpha_1)$$

$$\frac{Y^+}{S_1}\widetilde{tp} \in tp \pm Y^+\cdot(3\alpha_1)$$

Since, $tp \geq \frac{\gamma\cdot Y^+}{(2-\gamma)}$, we have $Y^+ \leq \frac{tp\cdot(2-\gamma)}{\gamma}$.

$$\frac{Y^+}{S_1}\widetilde{tp} \in tp \pm (3\alpha_1)\cdot\frac{tp\cdot(2-\gamma)}{\gamma}$$

$$\frac{Y^+}{S_1}\widetilde{tp} \in \left(1 \pm (3\alpha_1)\cdot\frac{(2-\gamma)}{\gamma}\right)tp$$

$$\widetilde{tp} \in (1\pm\epsilon)\,tp\frac{S_1}{Y^+}$$

where, $\epsilon = (3\alpha_1)\cdot\frac{(2-\gamma)}{\gamma}$

Therefore, $\alpha_1 = \frac{\gamma\cdot\epsilon}{3(2-\gamma)}$.

Now, size of samples required to satisfy the above approximation using Theorem 5.1 is, $S_1 = \Omega\left(\frac{1}{\alpha_1^2\cdot v}\left(d\log\frac{1}{v} + \log\frac{1}{\delta}\right)\right) = \Omega\left(\frac{(2-\gamma)^2}{\gamma^2\cdot\epsilon^2}\left(d + \log\frac{1}{\delta}\right)\right)$, with probability $1-\delta$ for all $q \in Q_\gamma \subset Q$.

For fn approximation, we can similarly show that

$$\widetilde{fn} \in fn\left(\frac{S_2}{Y^+}\right) \pm S_2\cdot(3\alpha_2).$$

Using $\epsilon = (3\alpha_2)$, we have the additive error to be ϵS_2.

Now, size of samples required to satisfy the above approximation using Theorem 5.1 is, $S_2 = \Omega\left(\frac{1}{\alpha_2^2\cdot v}\left(d\log\frac{1}{v} + \log\frac{1}{\delta}\right)\right) = \Omega\left(\frac{1}{\epsilon^2}\left(d + \log\frac{1}{\delta}\right)\right)$, with probability $1-\delta$ for all $q \in Q_\gamma \subset Q$.

For fp approximation,

$$\forall q \in Q; d_v\left(\sum_{z\in Y^-}\frac{1}{Y^-}\delta_z, \frac{1}{|S_3|}\sum_{z\in S_3}\delta_z\right) \leq \alpha_3$$

where, $\delta_z = 1(z \in fp)$ and $d_v(a,b) = \frac{|a-b|}{a+b+v}$. Again, using similar arguments, we can show that

$$\widetilde{fp} \in fp\left(\frac{S_3}{Y^-}\right) \pm S_3 \cdot \epsilon,$$

where again $\epsilon \geq 3\alpha_3$.

For $\widetilde{\widetilde{fp}}$, we get the following bound.

$$\widetilde{\widetilde{fp}} = \left(\frac{Y^-}{Y^+}\right)\widetilde{fp} \in fp\left(\frac{S_3}{Y^+}\right) \pm S_3 \cdot (3\alpha_3)\left(\frac{Y^-}{Y^+}\right)$$

After applying reweighing our modified α_3 is defined as, $\alpha_3^{'} = \frac{\epsilon}{3}\left(\frac{Y^+}{Y^-}\right)$.

Now, size of samples require to satisfy the above approximation using Theorem 5.1 is, $S_3 = \Omega\left(\frac{1}{(\alpha_3^{'})^2 \cdot v}\left(d\log\frac{1}{v} + \log\frac{1}{\delta}\right)\right) = \Omega\left(\left(\frac{Y^-}{Y^+}\right)^2 \frac{1}{\epsilon^2}\left(d + \log\frac{1}{\delta}\right)\right)$, with probability $1-\delta$ for all $q \in Q$ where $Q_\gamma \subset Q$.

Thus, in-order to satisfy all three approximations viz. tp, fn and fp we would require total sample size,

$$\begin{aligned} S &= S_1 + S_2 + S_3 \\ &= \left(\frac{(2-\gamma)^2}{\gamma^2\epsilon^2} + \frac{1}{\epsilon^2} + \left(\frac{Y^-}{Y^+}\right)^2 \frac{1}{\epsilon^2}\right) \cdot \left(d + \log\frac{1}{\delta}\right) \end{aligned}$$

Thus, with probability $1-3\delta$, above approximation for tp, fn and fp holds.

Now, for the coreset left hand side guarantee,

$$\begin{aligned} \tilde{F_1} &= \frac{\tilde{tp}}{\tilde{tp} + \frac{1}{2}(\tilde{fn} + \widetilde{\widetilde{fp}})} \\ &\geq \frac{(1-\epsilon)\frac{S\cdot tp}{Y^+}}{(1-\epsilon)\frac{S\cdot tp}{Y^+} + \frac{1}{2}\left[\frac{S\cdot fn}{Y^+} + \frac{S\cdot fp}{Y^+} + (\epsilon\cdot S)\left(1+\frac{Y^-}{Y^+}\right)\right]} \\ &= \frac{(1-\epsilon)tp}{(1-\epsilon)tp + \frac{1}{2}\left[fn + fp + (\epsilon)\left(Y^+ + Y^-\right)\right]} \\ &= \frac{(1-\epsilon)tp}{(1-\epsilon)tp + \frac{1}{2}\left[fn + fp + (\epsilon\cdot n)\right]} \\ &= \frac{(1-\epsilon)tp}{tp + \frac{1}{2}(fn+fp) - \left[\epsilon\cdot tp - \frac{\epsilon\cdot n}{2}\right]} \end{aligned}$$

If the number of samples belongs to tp is chosen to satisfy $tp \geq \frac{n(1-c\cdot\epsilon)}{2c\cdot(1-\epsilon)}$, then we have that

$$\begin{aligned} \tilde{F_1} &\geq \frac{(1-\epsilon)tp}{tp + \frac{1}{2}(fn+fp) - \left[\epsilon\cdot tp - \frac{\epsilon\cdot n}{2}\right]} \\ &\geq \frac{(1-c\cdot\epsilon)tp}{tp + \frac{1}{2}(fn+fp)} = (1-c\cdot\epsilon)F_1 \end{aligned}$$

Similarly, for the coreset right hand side guarantee,

$$\begin{aligned} \tilde{F_1} &= \frac{\tilde{tp}}{\tilde{tp} + \frac{1}{2}(\tilde{fn} + \widetilde{\widetilde{fp}})} \\ &\leq \frac{(1+\epsilon)\frac{S\cdot tp}{Y^+}}{(1+\epsilon)\frac{S\cdot tp}{Y^+} + \frac{1}{2}\left[\frac{S\cdot fn}{Y^+} + \frac{S\cdot fp}{Y^+} - (\epsilon\cdot S)\left(1+\frac{Y^-}{Y^+}\right)\right]} \\ &= \frac{(1+\epsilon)tp}{(1+\epsilon)tp + \frac{1}{2}\left[fn + fp - (\epsilon)\left(Y^+ + Y^-\right)\right]} \\ &= \frac{(1+\epsilon)tp}{(1+\epsilon)tp + \frac{1}{2}\left[fn + fp - (\epsilon\cdot n)\right]} \\ &= \frac{(1+\epsilon)tp}{tp + \frac{1}{2}(fn+fp) + \left[\epsilon\cdot tp - \frac{\epsilon\cdot n}{2}\right]} \end{aligned}$$

Again, using the similar assumptions on tp, we have that for $tp \geq \frac{n(1+c\cdot\epsilon)}{2c\cdot(1+\epsilon)}$, we get

$$\begin{aligned} \tilde{F_1} &\leq \frac{(1+\epsilon)tp}{tp + \frac{1}{2}(fn+fp) + \left[\epsilon\cdot tp - \frac{\epsilon\cdot n}{2}\right]} \\ &\leq \frac{(1+c\cdot\epsilon)tp}{tp + \frac{1}{2}(fn+fp)} = (1+c\cdot\epsilon)F_1 \end{aligned}$$

□

Weak Coreset for MCC

Theorem 5. *Let $\epsilon > 0$. Consider an instance where number of positive samples are Y^+ and number of negative samples are Y^-, and $n = Y^+ + Y^-$. Let T to be the ground truth positive $0/1$ labels and P to be the predicted positive $0/1$ labels. Let tp, fp, fn be the true positive, false positive and false negative on the full data, $T^{'} = \frac{\sum_i T_i}{n} = \frac{tp+fn}{n} = \frac{|Y^+|}{n}$ and $P^{'} = \frac{\sum_i P_i}{n} = \frac{tp+fp}{n}$. We consider Q_γ to be the set of queries such that $tp \geq \gamma\cdot n$ and $tn \geq \gamma\cdot n$ for $q \in Q_\gamma$. Let $d = vc-dimension\,(Q_\gamma)$. We claim that uniform sampling with $\left(\frac{1}{\epsilon^2}\left(d+\log\frac{1}{\delta}\right)\left(2+2\cdot\left(\frac{Y^-}{Y^+}\right)^2\right)\right)$ samples would be able to give a coreset for Q_γ that satisfies $\frac{MCC(q)}{\left(1+\frac{\epsilon}{\gamma}\right)} - 2\cdot\epsilon\cdot C \leq \widetilde{MCC}(q) \leq \frac{MCC(q)}{\left(1-\frac{\epsilon}{\gamma}\right)} + 2\cdot\epsilon\cdot C^{'}$ for all queries $q \in Q_\gamma$ with probability at least $1-4\delta$, where $C \leq \frac{\left(\frac{1}{\gamma}\right)}{\left(1+\frac{\epsilon}{\gamma}\right)\sqrt{T^{'}(1-T^{'})\cdot\gamma}}$ and $C^{'} \leq \frac{\left(\frac{1}{\gamma}\right)}{\left(1-\frac{\epsilon}{\gamma}\right)\sqrt{T^{'}(1-T^{'})\cdot\gamma}}$.*

The proof can be found in the full arxiv version.

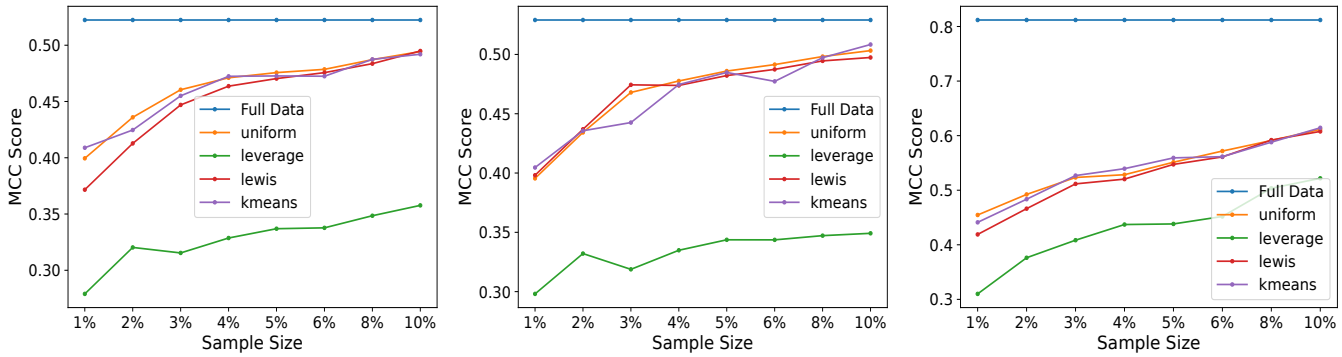
Experiments

All experiments were run on a computer with Nvidia Tesla V100 GPU with 32 GB memory and 28 CPUs. We used Python and its frameworks to implement our experiments.

Data Sets: The COVERTYPE (Blackard 1998) data consists of 581, 012 cartographic observations of different forests with 54 features. The task is to predict the type of trees at each location (49% positive). We selected in a stratified way

Coreset Algorithm	CoverType Time (in sec)	Adult Time (in sec)	KDD Cup '99 Time (in sec)
uniform	0.01418	0.003462	0.01028
leverage score	0.1331	0.01132	0.0909
lewis score	2.4918	0.19029	1.9072
k-means	0.05942	0.007611	0.0400

Table 1: Time taken to prepare coreset of 10% of full dataset for different datasets.



(a) CovType Logistic (b) CovType SVM (c) CovType MLP

Figure 1: MCC Score on CovType dataset for different classifiers and different coreset strategies.

50,000 samples from the real data and used them as our training data. The KDDCUP '99 (Stolfo et al. 1999) data comprises of 494,021 network connections with 41 features, and the task is to detect network intrusions (20% positive). In our experiments, we selected in a stratified way 50,000 samples from the real data and used them as training data. The Adult (Becker and Kohavi 1996) dataset is a widely-used dataset containing information about individuals from the 1994 U.S. Census Bureau database. It consists of approximately 32,000 instances with 14 attributes, including age, education, and occupation. The dataset aims to predict whether an individual's income exceeds $50,000 per year.

Our experiment created binary classification datasets by converting a multiclass dataset into a binary by flipping labels. Our final binary dataset class ratios for the CoverType dataset are [31770, 18230], for the Adult dataset [24294, 7706], and for the KDDCup dataset, it is [40154, 9846].

Experimental Assessment : To verify our theoretical claims, we tested a uniform sampling coreset with some of the sophisticated coresets like leverage score (Drineas, Mahoney, and Muthukrishnan 2006), l_1-lewis score (Cohen and Peng 2015), and k-means coreset by (Bachem, Lucic, and Krause 2018).

As our classifier models, we used SVM classifier with linear kernel; vanilla logistic regression model from the sklearn library is used with default hyperparameters and a multilayer perceptron(MLP). For MLP experiments, we considered a simple MLP classifier with two hidden layers of size 100 each and the final output layer of size two, as we are dealing with binary classification. The optimizer used for the MLP is Adam, and the activation function used is ReLU.

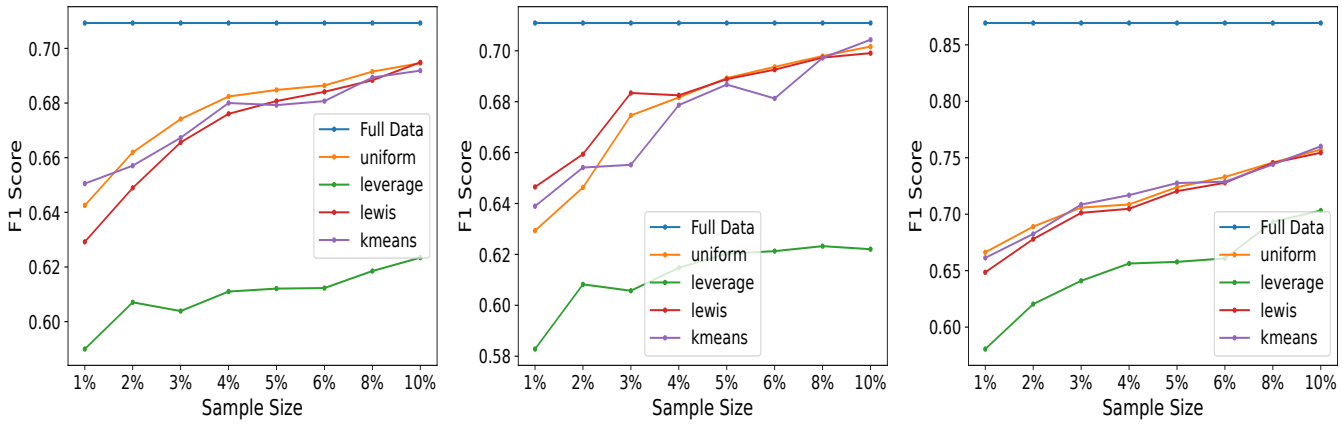
We first prepared different coresets from the full datasets. Then we train our classifier models: logistic regression, SVM, and Feed Forward Neural Network on our coreset as well as on the full dataset.

For testing the performance of our coresets, we took models trained using the coresets and tested them on the entire training dataset. We report the evaluation measures (F1 and MCC) obtained and compare it with the original value obtained on full data(using model trained on full data). For all our experiments, plotted values are means taken over five independent repetitions of each experiment.

Figures 1 through 4 clearly show that uniform sampling gives superior or comparable performance to other sophisticated methods for both F1 score and MCC. Also it can be seen that with increasing coreset size, the performance of model trained on the coreset also improves as expected. We also measure the time required to prepare coreset using these techniques on the different datasets. The times are reported in Table 1. It is clear that uniform sampling is many times faster than the other methods. Hence at much lower computation times we get better or comparable performances to other coreset construction strategies. Some additional experiments can be found in the appendix of full arxiv version.

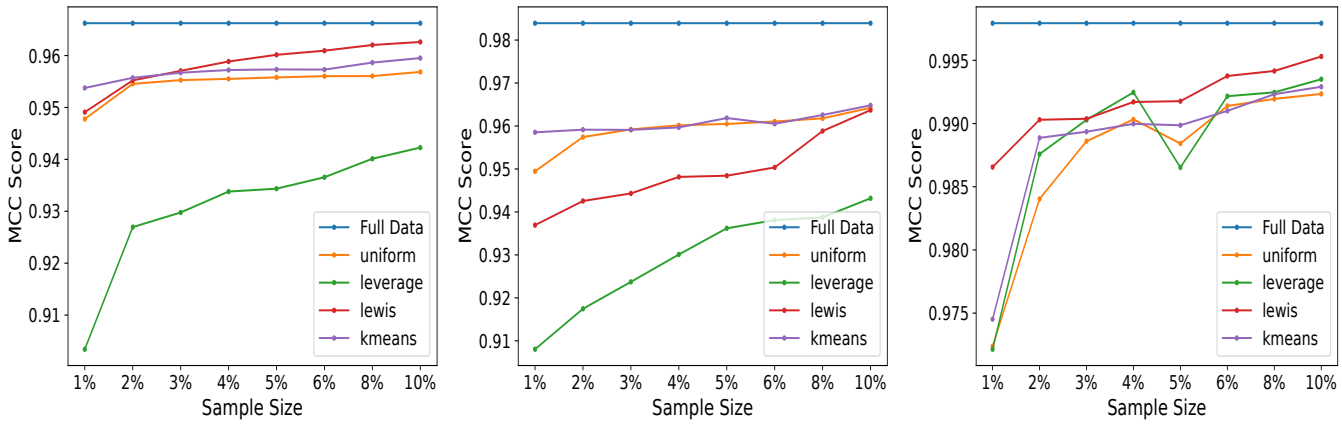
Conclusion

We initiated the study of coresets for non-decomposable classification measures, specifically for the F1 score and MCC. We showed lower bounds for strong coresets and construction of weak coresets using stratified uniform sampling. It would be interesting to see whether coresets with better additive guarantees and lesser assumptions on the query vector can be developed. Similarly, algorithm-specific subset selection strategies could be explored for more efficiency. The question of tackling other measures, e.g., AUC-ROC, remains open.



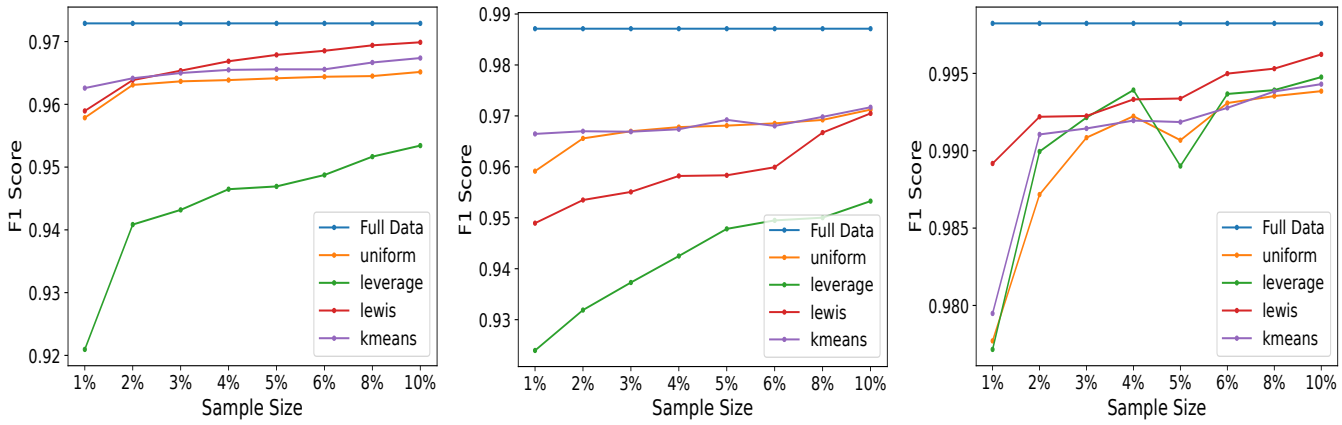
(a) CovType Logistic (b) CovType SVM (c) CovType MLP

Figure 2: F_1 Score on CovType dataset for different classifiers and different coreset strategies.



(a) KDDCup Logistic (b) KDDCup SVM MCC (c) KDDCup MLP MCC

Figure 3: MCC Score on KDDCup dataset for different classifiers and different coreset strategies.



(a) KDDCup Logistic (b) KDDCup SVM (c) KDDCup MLP

Figure 4: F_1 Score on KDDCup dataset for different classifiers and different coreset strategies.

Acknowledgments

Anirban Dasgupta would like to acknowledge the support received from Google, DST(SERB) and Cisco as well as the N Rama Rao Chair position at IIT Gandhinagar. Jayesh Malaviya and Rachit Chhaya would like to acknowledge the support from IIT- Gandhinagar and Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT) , Gandhinagar, India.

References

Agarwal, P. K.; Har-Peled, S.; Varadarajan, K. R.; et al. 2005. Geometric approximation via coresets. *Combinatorial and computational geometry*, 52(1): 1–30.

Bachem, O.; Lucic, M.; and Krause, A. 2017. Practical coreset constructions for machine learning. *arXiv preprint arXiv:1703.06476*.

Bachem, O.; Lucic, M.; and Krause, A. 2018. Scalable k-means clustering via lightweight coresets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1119–1127.

Becker, B.; and Kohavi, R. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Bénédict, G.; Koops, H. V.; Odijk, D.; and de Rijke, M. 2022. sigmoidF1: A Smooth F1 Score Surrogate Loss for Multilabel Classification. *Transactions on Machine Learning Research*.

Blackard, J. 1998. Covertype. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C50K5N.

Braverman, V.; Cohen-Addad, V.; Jiang, H.-C. S.; Krauthgamer, R.; Schwiegelshohn, C.; Toftrup, M. B.; and Wu, X. 2022. The power of uniform sampling for coresets. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, 462–473. IEEE.

Braverman, V.; Feldman, D.; Lang, H.; Statman, A.; and Zhou, S. 2016. New frameworks for offline and streaming coreset constructions. *arXiv preprint arXiv:1612.00889*.

Cohen, M. B.; and Peng, R. 2015. Lp row sampling by lewis weights. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, 183–192.

Drineas, P.; Mahoney, M. W.; and Muthukrishnan, S. 2006. Sampling algorithms for l 2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, 1127–1136.

Eban, E.; Schain, M.; Mackey, A.; Gordon, A.; Rifkin, R.; and Elidan, G. 2017. Scalable learning of non-decomposable objectives. In *Artificial intelligence and statistics*, 832–840. PMLR.

Feldman, D. 2020. Core-sets: Updated survey. *Sampling Techniques for Supervised or Unsupervised Tasks*, 23–44.

Feldman, D.; and Langberg, M. 2011. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, 569–578.

Joachims, T. 2005. A support vector method for multivariate performance measures. In *Proceedings of the 22nd international conference on Machine learning*, 377–384.

Kar, P.; Narasimhan, H.; and Jain, P. 2014. Online and stochastic gradient methods for non-decomposable loss functions. *Advances in Neural Information Processing Systems*, 27.

Langberg, M.; and Schulman, L. J. 2010. Universal ε-approximators for integrals. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, 598–607. SIAM.

Li, Y.; Long, P. M.; and Srinivasan, A. 2001. Improved bounds on the sample complexity of learning. *Journal of Computer and System Sciences*, 62(3): 516–527.

Lu, F.; Raff, E.; and Holt, J. 2023. A Coreset Learning Reality Check. *arXiv preprint arXiv:2301.06163*.

Mai, T.; Musco, C.; and Rao, A. 2021. Coresets for classification–simplified and strengthened. *Advances in Neural Information Processing Systems*, 34: 11643–11654.

Munteanu, A.; Schwiegelshohn, C.; Sohler, C.; and Woodruff, D. 2018. On coresets for logistic regression. *Advances in Neural Information Processing Systems*, 31.

Nan, Y.; Chai, K. M.; Lee, W. S.; and Chieu, H. L. 2012. Optimizing F-measure: A tale of two approaches. *arXiv preprint arXiv:1206.4625*.

Narasimhan, H.; Kar, P.; and Jain, P. 2015. Optimizing non-decomposable performance measures: A tale of two classes. In *International Conference on Machine Learning*, 199–208. PMLR.

Poms, F.; Sarukkai, V.; Mullapudi, R. T.; Sohoni, N. S.; Mark, W. R.; Ramanan, D.; and Fatahalian, K. 2021. Low-shot validation: Active importance sampling for estimating classifier performance on rare categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10705–10714.

Samadian, A.; Pruhs, K.; Moseley, B.; Im, S.; and Curtin, R. 2020. Unconditional coresets for regularized loss minimization. In *International Conference on Artificial Intelligence and Statistics*, 482–492. PMLR.

Sanyal, A.; Kumar, P.; Kar, P.; Chawla, S.; and Sebastiani, F. 2018. Optimizing non-decomposable measures with deep networks. *Machine Learning*, 107: 1597–1620.

Sawade, C.; Landwehr, N.; and Scheffer, T. 2010. Active estimation of f-measures. *Advances in Neural Information Processing Systems*, 23.

Stolfo, S.; Fan, W.; Lee, W.; Prodromidis, A.; and Chan, P. 1999. KDD Cup 1999 Data. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C51C7N.

Tukan, M.; Baykal, C.; Feldman, D.; and Rus, D. 2021. On coresets for support vector machines. *Theoretical Computer Science*, 890: 171–191.