Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 195–205
DOI: 10.15439/2025F0110
ISSN 2300-5963 ACSIS, Vol. 43

Estimating the entropy of covering-based rough set approximation operators

Wannes De Maeyer, Steven Van Overberghe, Chris Cornelis
0009-0005-1597-1023
0000-0002-0818-468X
0000-0002-6852-4041
Department of Mathematics, Computer Science and Statistics
Ghent University,
Ghent, Belgium
Email: {wannes.demaeyer, steven.vanoverberghe, chris.cornelis}@ugent.be

Mauricio Restrepo
0000-0003-1007-353X
Department of Mathematics
Universidad Militar Nueva Granada,
Bogotá, Colombia
Email: mauricio.restrepo@unimilitar.edu.co

***Abstract*—Rough set theory provides a robust framework for dealing with inconsistent data by utilizing equivalence relations to group indiscernible instances. A significant extension of this framework is the concept of covering-based rough sets, where equivalence relations are replaced with coverings. The effectiveness of covering-based rough sets in practical applications largely depends on the choice of an appropriate covering. To guide this selection, various metrics have been proposed to evaluate the quality of coverings, including entropy. While entropy serves as a valuable measure, its exact computation is often prohibitively expensive. In this paper, we propose an efficient method for estimating the entropy of covering-based rough set approximation operators, making the metric more feasible for practical use. We assess the accuracy of these estimates through experiments on both synthetic coverings and coverings for real-world datasets, the latter constructed using the Mapper algorithm, from the field of topological data analysis.**

I. Introduction

IN 1982 Pawlak introduced rough set theory as a framework to handle possibly inconsistent data [1]. Rough set theory revolves around the notion of an information system: a couple $(U, \mathcal{A})$, where U is a universe of instances (objects, data samples) and $\mathcal{A}$ is a set of attributes for which these instances take values. The set $\mathcal{A}$ naturally gives rise to an equivalence relation E on U, by partitioning U into equivalence classes that have the same values for all the attributes in $\mathcal{A}$. Instances in each equivalence class are called indiscernible. Rough set theory then proceeds to provide lower and upper approximations of concepts that are represented by a set $A \subseteq U$. The lower approximation is the union of all equivalence classes that are contained entirely within A. This can be viewed as the set of instances that certainly belong to A (every instance in the lower approximation is only indiscernible to elements of A). The upper approximation is the union of all equivalence classes that have non-empty intersection with A. Semantically this is the set of instances that possibly belong to U (every instance in the upper approximation is indiscernible with an element in A). Sets $A \subseteq U$ for which the lower approximation and upper approximation equal A itself are called consistent, or exact. It can be verified that the lower approximation is the largest consistent subset of A, while the upper approximation is the smallest consistent set that includes A.

Various researchers have proposed generalizations of classical rough sets by replacing the equivalence relation with a covering of U [2, 3, 4], resulting in a more flexible framework. Recall that a covering of U is any collection of subsets of U whose union is equal to U. By contrast to a partition, elements of a covering are allowed to overlap. This is useful in various situations, for example:

- handling missing data [5]: when the value of an instance for an attribute is unknown, it is considered indiscernible from any other (known or unknown) value.
- tolerance-based rough sets [6]: when dealing with numerical data, two objects are often considered indiscernible when their distance is lower than a given threshold.

Clearly, such indiscernibility relations are no longer transitive. These covering-based rough sets have been studied extensively from a theoretical perspective ([7, 8, 9, 10]) and used in many applications, in particular concerning attribute reduction [11, 12, 13]. However, selecting the right covering for an application can be hard because there may be multiple important factors to consider. One such factor is the granularity of the covering. This measures how fine the covering is and how detailed the rough set approximations are; it can also be interpreted as the amount of information that remains when the approximation operators are executed. To find coverings that have a certain granularity, the concept of entropy (which was originally introduced as a measure of information content) was extended to rough set approximations in [14]. Many applications in rough set theory use the notion of entropy [15, 16]. However, when it comes to covering-based rough set theory, it is computationally very expensive to calculate this naively since it takes into account all of the possible subsets of the universe U. For this reason, we propose an efficient way to estimate the entropy.

In [17] a promising method to generate coverings of datasets using the Mapper algorithm from the field of topological data analysis [18] was introduced. We use these coverings

to evaluate the convergence of our estimates. We choose this method because it produces coverings that closely resemble the original dataset topologically. Moreover, one can easily control several aspects of the coverings, such as the number of subsets in the covering, and the amount of overlap between them.

The remainder of this paper is structured as follows. We first recall the concepts of covering-based rough sets and entropy in Sections 2 and 3. Then, in Section 4, we outline a method to estimate the entropy of a given covering using the so-called strong granule-based approximations from covering-based rough set theory. To obtain the estimation, we first consider the method from [19] which allows us to estimate the entropy of any probability distribution. However, we observe that it is restrictively slow for our purposes. Therefore, we adapt this method using some properties of covering-based rough sets to find better estimates that are calculated more quickly. In Section 5, we evaluate the convergence of our estimates experimentally on a number of synthetic and benchmark datasets. Finally, in Section 6, we give a discussion and introduce some ideas for future exploration.

II. Covering-based rough sets

When working with real-world classification problems, inconsistencies often occur in datasets; intuitively, this happens when two instances are indiscernible but belong to different classes. Pawlak introduced rough sets in 1982 to handle such inconsistencies [1]. In Pawlak's seminal work, an approximation space is defined as an ordered pair (U, E), where E is an equivalence relation that groups indiscernible instances together over the universe U. Based on this setup, the lower and upper approximations of a set can be described in several equivalent ways. In this paper, we adopt the granule-based approach:

$$\underline{apr}(A) = \{[x]_E \in U/E \mid [x]_E \subseteq A\}$$
$$\overline{apr}(A) = \{[x]_E \in U/E \mid [x]_E \cap A \neq \emptyset\}$$

For a subset A of U, its rough set approximation is represented as the pair $(\underline{apr}(A), \overline{apr}(A))$. These approximation operators are dual in the sense that $\underline{apr}(A) = \overline{apr}(A^c)^c$, where ^{c} refers to the classical set complement. These approximations have been used in many applications ranging from classification [20] (especially rule extraction [21, 22]) to feature selection [23, 24, 25].

Classical rough set theory often struggles with continuous features, as the indiscernibility relation may no longer satisfy the properties of an equivalence relation. To overcome this limitation, various generalizations of classical rough sets have been proposed. One such generalization are covering-based rough sets, which replace the partition given by the indiscernibility relation with a covering. This defines a covering approximation space: an ordered pair $(U, \mathbb{C})$, where $\mathbb{C}$ is a covering of U, that is, a collection of subsets whose union equals U. While it is tempting to generalize Pawlak's definitions by replacing the equivalence classes $[x]_E \in U/E$ with elements $K \in \mathbb{C}$, this substitution does not preserve the duality between the approximations. To address this, Yao [2] introduced two dual pairs of granule-based approximations for covering approximation spaces.

$$\underline{apr'}_{\mathbb{C}}(A) = \bigcup\{K \in \mathbb{C} : K \subseteq A\} \tag{1}$$
$$\overline{apr'}_{\mathbb{C}}(A) = \underline{apr'}_{\mathbb{C}}(A^c)^c \tag{2}$$
$$\underline{apr''}_{\mathbb{C}}(A) = \overline{apr''}_{\mathbb{C}}(A^c)^c \tag{3}$$
$$\overline{apr''}_{\mathbb{C}}(A) = \bigcup\{K \in \mathbb{C} \mid K \cap A \neq \emptyset\} \tag{4}$$

Equations (1) and (2) are called the strong approximation operators, while Equations (3) and (4) are referred to as the weak approximations. The rationale behind this terminology is explained by the following property:

$$\underline{apr''}_{\mathbb{C}}(A) \subseteq \underline{apr'}_{\mathbb{C}}(A) \subseteq A \subseteq \overline{apr'}_{\mathbb{C}}(A) \subseteq \overline{apr''}_{\mathbb{C}}(A)$$

Example 1. *Let $U = \{x_1, \ldots, x_9\}$ be a universe which has the following covering, $\mathbb{C} = \{\{x_0, x_1\}, \{x_1, x_2\}, \{x_2, x_5\}, \{x_3, x_4, x_7\}, \{x_6, x_7, x_8, x_9\}\}$ and $A = \{x_1, x_2, x_3, x_4, x_5\}$. A has the following approximations:*

$$\underline{apr'}_{\mathbb{C}}(A) = \{x_1, x_2, x_5\}$$
$$\overline{apr'}_{\mathbb{C}}(A) = \{x_0, x_1, x_2, x_3, x_4, x_5\}$$
$$\underline{apr''}_{\mathbb{C}}(A) = \{x_2, x_5\}$$
$$\overline{apr''}_{\mathbb{C}}(A) = \{x_0, x_1, x_2, x_3, x_4, x_5, x_7\}$$

Generating suitable coverings for a particular application is an important but challenging task. To assist with this process, we associate an ordering to the set of all coverings.

Definition 1. *[9] Let $\mathbb{C}$ and $\mathbb{C}'$ be two coverings of U. $\mathbb{C} \ll \mathbb{C}'$ if and only if for all $L \in \mathbb{C}'$ there exists a set $S \subseteq \mathbb{C}$ such that $L = \cup_{K \in S} K$. When $\mathbb{C} \ll \mathbb{C}'$, we also say that $\mathbb{C}$ is finer than $\mathbb{C}'$.*

We have the following proposition [9]:

Proposition 1. *Let $\mathbb{C}$ and $\mathbb{C}'$ be coverings of U. Then $\underline{apr'}_{\mathbb{C}'}(A) \subseteq \underline{apr'}_{\mathbb{C}}(A) \subseteq A \subseteq \overline{apr'}_{\mathbb{C}}(A) \subseteq \overline{apr'}_{\mathbb{C}'}(A)$ for all $A \subseteq U$, if and only if $\mathbb{C} \ll \mathbb{C}'$.*

Example 2. *To illustrate the previous proposition we continue Example 1. Let $\mathbb{C}' = \{\{x_0, x_1, x_2\}, \{x_2, x_5\}, \{x_3, x_4, x_6, x_7, x_8, x_9\}\}$. Then we clearly have that $\mathbb{C} \ll \mathbb{C}'$. If we now recalculate the approximations of A, using $\mathbb{C}$, we get:*

$$\underline{apr'}_{\mathbb{C}'}(A) = \{x_2, x_5\}$$
$$\overline{apr'}_{\mathbb{C}'}(A) = \{x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9\}$$

In [17] a promising new method was introduced to generate coverings for classification problems, using the Mapper algorithm that originated in the field of topological data analysis (TDA). It generates coverings of a universe U, as shown in Algorithm 1, which generally represent the topological structure of the dataset and are therefore useful for our purposes. The algorithm uses a lens function $f : U \to \mathbb{R}^d$, an input covering $\mathbb{I}$ of $f(U)$, and a clustering algorithm. Often

the lens function is chosen to be a projection on particular attributes or PCA components. The input covering $\mathbb{I}$ is in practice constructed using a number of (hyper-)cubes among each dimension in $\mathbb{R}^d$, such that consecutive cubes have a certain overlap percentage p.

Mapper proceeds in the following way: the lens function transforms the original dataset U to the lower-dimensional space $f(U)$, on which the covering $\mathbb{I}$ is defined. Then, for each I inside the input covering, the instances from U projected onto it by f (i.e., $f^{-1}(I)$) are partitioned by means of a clustering algorithm, another parameter of the algorithm. These clusters form the resulting covering.

In [26] we proved the following result which holds when the clustering algorithm is equal to the single-linkage clustering algorithm [27] (shown in Algorithm 2).

Proposition 2. *Let U be a dataset, $f : U \to \mathbb{R}^n$. Let K denote the single linkage clustering algorithm. Let $\mathbb{I}$ and $\mathbb{I}'$ be coverings of $f(U)$, and let $\mathbb{C}_{\mathbb{I},\varepsilon}$ and $\mathbb{C}_{\mathbb{I}',\varepsilon'}$ denote the coverings generated by the Mapper algorithm using lens function f, the single linkage algorithm with parameters ε and ε' and coverings $\mathbb{I}$ and $\mathbb{I}'$. If $\mathbb{I} \ll \mathbb{I}'$ and $\varepsilon \leq \varepsilon'$, then $\mathbb{C}_{\mathbb{I},\varepsilon} \ll \mathbb{C}_{\mathbb{I}',\varepsilon'}$.*

Algorithm 1: The Mapper algorithm

Input: a dataset U with an associated metric, a *lens* function $f : U \to \mathbb{R}$ (or $\mathbb{R}^m$), a covering $\mathbb{I} = \{I_j \mid j = 1, \dots n\}$ of $f(U)$, and a clustering algorithm cl on U.

for $j \in \{1, 2, \dots, n\}$ **do**

  $A_j = \{A_{j1}, \dots, A_{jk} \mid i = 1, \dots k\} \leftarrow cl(f^{-1}(I_j))$;

end

Result: $\cup A_j$

Algorithm 2: The Single Linkage algorithm

Input: a dataset U, with an associated metric d and a real number $\varepsilon > 0$

Construct a graph G_ε, with U as vertices and an edge between x and y when $d(x, y) \leq \varepsilon$. ;

$\{A_1, \dots, A_k\} \leftarrow$ The connected components of G_ε;

Result: $\{A_1, \dots, A_k\}$

III. Entropy

In order to be able to control the fineness of the constructed coverings, we require a measure that reflects granularity. A suitable candidate for this is entropy, which is classically defined for probability distributions:

Definition 2. *[28] Let $\Pr$ be a probability distribution over a finite universe U. The entropy of $\Pr$ is defined as:*

$$H(\Pr) := -\sum_{x \in U} \Pr(x) \log \Pr(x)$$

This definition can be naturally adjusted to partitions:

Definition 3. *[19] Let $\pi = \{A_1, \dots, A_k\}$ be a partition of U. We define the entropy of π as:*

$$H(\pi) := -\sum_{i=1}^{k} \frac{|A_i|}{|U|} \log \frac{|A_i|}{|U|}.$$

The entropy indeed reflects the fineness of a partition. When every element resides in its own equivalence class, the entropy is maximal and equal to $\log |U|$. When there is only one equivalence class, the entropy is minimal and equal to 0.

We also have the following property:

Proposition 3. *[14] Let $\pi \ll \pi'$ be two partitions of U, then $H(\pi) \geq H(\pi')$.*

Since we work with coverings instead of partitions, Definition 3 is not directly applicable. Instead, we adopt the proposal from [14] that computes the entropy associated with an arbitrary pair of covering-based approximation operators.

Definition 4. *[14]*

- *Let $(\underline{apr}, \overline{apr})$ be a pair of covering-based approximation operators over a finite universe U. We define $\Pi_{\underline{apr}}^{\overline{apr}}$ as the set of equivalence classes of $\mathcal{P}(U)$ of the equivalence relation $A \approx B \Leftrightarrow (\underline{apr}(A), \overline{apr}(A)) = (\underline{apr}(B), \overline{apr}(B))$.*
- *We define the entropy of $(\underline{apr}, \overline{apr})$ as follows:*

$$\mathcal{H}_{\underline{apr}}^{\overline{apr}} = H(\Pi_{\underline{apr}}^{\overline{apr}})$$

Again we observe that a high entropy occurs when approximation operators are very fine and a low entropy occurs when approximation operators are very rough. When we have for all $A \subseteq U$ that $(\underline{apr}(A), \overline{apr}(A)) = (A, A)$ (the approximation operator is maximally fine), the size of each class in $\Pi_{\underline{apr}}^{\overline{apr}}$ is equal to 1 and thus the entropy reaches its the maximal value, $|U|$. When an approximation operator is maximally rough, i.e. $(\underline{apr}(A), \overline{apr}(A)) = (\emptyset, U)$ for all A, $\Pi_{\underline{apr}}^{\overline{apr}}$ is equal to $\{U\}$ and thus the entropy is minimal and equal to 0.

Example 3. *Let $U = \{x_1, \dots, x_n\}$ and $\mathbb{C} = \{\{x_i, x_{i+1}\} \mid 1 \leq i < n\}$.*

We calculate the entropy of the strong approximations by first constructing $\Pi_{\underline{apr'}_{\mathbb{C}}}^{\overline{apr'}_{\mathbb{C}}}$.

Let $\approx$ be the equivalence relation corresponding to $\Pi_{\underline{apr'}_{\mathbb{C}}}^{\overline{apr'}_{\mathbb{C}}}$. Then $A \approx B$ if either $A = B$ or A and B are equal to $\{x_i \mid 1 \leq i \leq n$ and $i \equiv 0$ (mod 2)$\}$ and $\{x_i \mid 1 \leq i \leq n$ and $i \equiv 1$ (mod 2)$\}$.

To see this, first note that

$$\underline{apr'}_{\mathbb{C}}(A) = \{x_i \mid x_i \in A \text{ and } (x_{i-1} \in A \text{ or } x_{i+1} \in A)\}$$
$$\overline{apr'}_{\mathbb{C}}(A) = \{x_i \mid x_i \in A \text{ or } (x_{i-1} \in A \text{ and } x_{i+1} \in A)\}$$

Suppose that $A \neq B$ but $A \approx B$. Let i be an integer such that $x_i \in A$ and $x_i \notin B$. Since $x_i \in \overline{apr'}_{\mathbb{C}}(A)$, $x_{i+1} \in B$ and $x_{i-1} \in B$. Since $x_i \notin \underline{apr'}_{\mathbb{C}}(B)$, $x_{i-1} \notin A$ and $x_{i+1} \notin A$. We can repeat the same procedure for x_{i-1} and x_{i+1} and in the end we conclude that A and B have the requested shape.

Thus, $\Pi_{\underline{apr'}_{\mathbb{C}}}^{\overline{apr'}_{\mathbb{C}}}$ has one equivalence class of size two and all other classes have size one. We can now calculate the entropy:

$$\begin{aligned}\mathcal{H}_{\underline{apr'}_{\mathbb{C}}}^{\overline{apr'}_{\mathbb{C}}} &= \frac{-\log 2^{-n+1}}{2^{n-1}} - \sum_{i=1}^{2^n-2} \frac{1}{2^n} \log \frac{1}{2^n} \\ &= \frac{n-1}{2^{n-1}} + \frac{(2^n-2)n}{2^n} \\ &= \frac{2n-2+n2^n-2n}{2^n} \\ &= n - \frac{1}{2^{n-1}}\end{aligned}$$

We also prove that entropy increases monotonically with increasing fineness in coverings. This is useful because, as mentioned in Proposition 2, we are able to construct chains of coverings under $\ll$. Therefore, we may exploit this monotonicity to guide the search for coverings that possess a certain entropy.

Proposition 4. *Let $\mathbb{C} \ll \mathbb{C}'$ be coverings of U. If $(\underline{apr'}_{\mathbb{C}}(A), \overline{apr'}_{\mathbb{C}}(A)) = (\underline{apr'}_{\mathbb{C}}(B), \overline{apr'}_{\mathbb{C}}(B))$ then also $(\underline{apr'}_{\mathbb{C}'}(A), \overline{apr'}_{\mathbb{C}'}(A)) = (\underline{apr'}_{\mathbb{C}'}(B), \overline{apr'}_{\mathbb{C}'}(B))$.*

Proof. First assume $\underline{apr'}_{\mathbb{C}}(A) = \underline{apr'}_{\mathbb{C}}(B)$. We now have for every $K \in \mathbb{C}$ that $K \subseteq A$ if and only if $K \subseteq B$. Assume by contradiction that $\underline{apr'}_{\mathbb{C}'}(A) \neq \underline{apr'}_{\mathbb{C}'}(B)$ then (w.l.o.g.) there exists a $K \in \mathbb{C}'$ such that $K \subseteq A$ but $K \nsubseteq B$. Because $\mathbb{C} \ll \mathbb{C}'$ we have that $K = \cup_{L \in S} L$ for some $S \subseteq \mathbb{C}$. However, for all $L \in S$, it holds that $L \subseteq A$ and thus $L \subseteq B$. This is a contradiction. Thus, $\underline{apr'}_{\mathbb{C}'}(A) = \underline{apr'}_{\mathbb{C}'}(B)$.

Because of the duality of the strong approximation operator, also $\overline{apr'}_{\mathbb{C}'}(A) = \overline{apr'}_{\mathbb{C}'}(B)$.

□

This implies the following theorem:

Theorem 1. *Let $\mathbb{C} \ll \mathbb{C}'$ be coverings of a universe U, then $\mathcal{H}_{\underline{apr'}_{\mathbb{C}}}^{\overline{apr'}_{\mathbb{C}}} \geq \mathcal{H}_{\underline{apr'}_{\mathbb{C}'}}^{\overline{apr'}_{\mathbb{C}'}}$.*

Proof. This follows from Proposition 4 and Proposition 3. □

There does not exist a similar result establishing monotonicity between the weak granule-based approximations and entropy. Because of this, we only consider the strong approximations in the remainder of this paper.

IV. Estimating entropy

Since our definition of entropy takes into account all subsets of U, it is computationally expensive to calculate its exact value. Because of this, we will estimate the value of the entropy. It is well-known that there is no unbiased estimator for the entropy of a probability distribution when there is no external knowledge about the probability distribution [29]. There exist, however, some good estimators that do not depend on external knowledge. We will also use our knowledge about covering-based rough sets to estimate and determine the exact size of certain equivalence classes of $\Pi_{\underline{apr}}^{\overline{apr}}$, which will lead to an unbiased estimator anyway.

We proceed in the following way: Let Y equal the multiset $\{-\frac{|\mathcal{A}_i|}{2^{|U|}} \log \frac{|\mathcal{A}_i|}{2^{|U|}} \mid \mathcal{A}_i \in \Pi_{\underline{apr}}^{\overline{apr}}\}$. Estimating the entropy is then equivalent to estimating the sum of Y.

Inspired by [19] and [30], we do this using the Horvitz–Thompson estimator [31]. This estimator is unbiased and counts the sum of all elements y_i in a universe Y by sampling a set S from Y without replacement. The estimator is equal to

$$\sum_{y_i \in S} \frac{y_i}{\Pr(y_i \in S)},$$

where $\Pr(y_i \in S)$ is the probability of a sample y_i belonging to S. However, this estimator is only unbiased if we know the size of each $\mathcal{A}_i$ exactly.

We proceed as follows: first, we generate n random subsets $\{B_1, \ldots, B_n\}$ of U. Let M be the set of approximations $\{(\underline{apr'}_{\mathbb{C}}(B_i), \overline{apr'}_{\mathbb{C}}(B_i)) \mid 1 \leq i \leq n\}$. We observe that the probability of B_i belonging to a certain equivalence class $\mathcal{A}_l$ of $\Pi_{\underline{apr}}^{\overline{apr}}$ is equal to the proportion of the size of the equivalence class to the number of possible subsets: $\frac{|\mathcal{A}_l|}{2^{|U|}}$.

We define $p_k = \frac{|\mathcal{A}_k|}{2^{|U|}}$ for every $1 \leq k \leq |M|$ (note that $|M| \leq n$), where $\mathcal{A}_k = \{B \subseteq U \mid (\underline{apr'}_{\mathbb{C}}(B), \overline{apr'}_{\mathbb{C}}(B)) = (\underline{A}_k, \overline{A}_k)\}$, and $(\underline{A}_k, \overline{A}_k)$ is the k-th element of M. The probability of not generating a set from $\mathcal{A}_k$ is equal to $(1-p_k)^n$ and thus the probability of generating a set inside $\mathcal{A}_k$ is equal to $1-(1-p_k)^n$. Therefore, the result from the Horvitz-Thomson estimator is:

$$\hat{\mathcal{H}}_{\underline{apr'}_{\mathbb{C}}}^{\overline{apr'}_{\mathbb{C}}} = -\sum_k \frac{p_k \log p_k}{1-(1-p_k)^n} \tag{5}$$

We also have an estimator for the variance of the Horvitz-Thomson estimator [31]:

$$\begin{aligned}&\hat{V}\left(\hat{\mathcal{H}}_{\underline{apr'}_{\mathbb{C}}}^{\overline{apr'}_{\mathbb{C}}}\right) \\ &= \left(\hat{\mathcal{H}}_{\underline{apr'}_{\mathbb{C}}}^{\overline{apr'}_{\mathbb{C}}}\right)^2 - \sum_k \frac{(p_k \log p_k)^2}{1-(1-p_k)^n} \\ &\quad - \sum_{k \neq j} \frac{(p_k \log p_k)(p_j \log p_j)}{1-(1-p_k)^n-(1-p_j)^n+(1-p_k-p_j)^n}\end{aligned} \tag{6}$$

When we have exact values for p_k available, both of these estimators are unbiased [31] and can thus be used to estimate the entropy and the corresponding confidence interval. A new problem presents itself: determining p_k. We can do this by empirically estimating p_k (Section IV-A) or we can use properties of the covering itself to determine it exactly (Section IV-B).

A. Coverage-adjusted estimator

The coverage-adjusted estimator was introduced in [19] and uses the Horvitz-Thomson estimator to estimate the entropy of a probability distribution. According to the observations of the authors of [19], the most intuitive way to estimate p_k empirically is by letting $\hat{p}_k$ equal the number of generated sets B_i that have $(\underline{A}_k, \overline{A}_k)$ as their approximations divided by n. However, they introduced a more useful approximation, where

the estimator converges faster to the true value of the entropy. Here $\hat{C} = 1 - \frac{f_1}{n}$, where f_1 equals the number of approximations $(\underline{A}_k, \overline{A}_k)$ such that there is only one B_i which has $(\underline{A}_k, \overline{A}_k)$ as its approximation. We will call approximations which are reached more than once collisions. We now define

$$\hat{p}_k = \frac{\hat{C}}{n} |\{B_i \mid (\underline{apr'}_{\mathbb{C}}(B_i), \overline{apr'}_{\mathbb{C}}(B_i)) = (\underline{A}_k, \overline{A}_k)\}| \quad (7)$$

This estimation of entropy based on Equation (5) and Equation (7) is called the coverage-adjusted estimator. It converges to the true value of entropy as follows. Let $\hat{\mathcal{H}}$ be the coverage-adjusted entropy estimate of $\mathcal{H}$ generated using n samples. We now have that $|\hat{\mathcal{H}} - \mathcal{H}| = O(1/\log(n))$ [19]

The main downside of this estimator is that we need to choose a large enough n to ensure that there is an approximation $(\underline{A}_k, \overline{A}_k)$ that is reached at least twice. When there are no such sets, $\hat{C} = 0$ and thus all $p_k = 0$ which will lead to an entropy of 0 (convention states that $0 \log 0 = 0$). When the entropy is high this often means that n needs to be unfeasibly large. When the entropy is equal to e, we expect this to happen on average after generating $O(2^{e/2})$ sets [32]. To mitigate this problem, the authors of [19] replaced the definition of $\hat{C}$ with $1 - \frac{f_1}{n+1}$. Now $\hat{C}$ is never equal to 0 and still has the same asymptotic behavior. However, when the true value of the entropy is too high, this can still cause problems. When no collisions occur the coverage-adjusted estimator is equal to:

$$\begin{aligned}
\hat{\mathcal{H}} &= -\sum_{i=1}^{n} \frac{\hat{C}/n \log(\hat{C}/n)}{1 - (1 - \hat{C}/n)^n} \\
&= \sum_{i=1}^{n} \frac{\log(n(n+1))}{n(n+1)(1 - (1 - \frac{1}{n(n+1)})^n)} \\
&= \frac{\log(n(n+1))}{(n+1)(1 - (\frac{n(n+1)-1}{n(n+1)})^n)} \\
&= O(\log(n))
\end{aligned}$$

We would thus need exponentially many samples to estimate an entropy accurately when no collisions occur. This can be intractable when the entropy is large.

B. Horvitz-Thompson Exact Estimator

In this section we introduce a different way of estimating the entropy by determining the exact size of $\mathcal{A}_k$ so that we can use Equation (5) to get an unbiased estimator.

To do this, we introduce the following definition: $P_{\mathbb{C}}$ is the partition of U given by the equivalence relation $\sim$, defined as follows:

$$(\forall x, y \in U)(x \sim y \Leftrightarrow \{K \in \mathbb{C} \mid x \in K\} = \{K \in \mathbb{C} \mid y \in K\})$$

In other words, points that are equivalent always occur together in the covering. Also note that every element K of $\mathbb{C}$ can be written as the union of elements in $P_{\mathbb{C}}$.

We will call elements of $P_{\mathbb{C}}$ blocks, and we define a tripartition of $P_{\mathbb{C}}$ for every $A \subseteq U$:

$$\begin{aligned}
P_1(A) &:= \{X \in P_{\mathbb{C}} \mid X \subseteq A\} \\
P_2(A) &:= \{X \in P_{\mathbb{C}} \mid \emptyset \neq X \cap A \neq X\} \\
P_3(A) &:= \{X \in P_{\mathbb{C}} \mid X \cap A = \emptyset\}
\end{aligned}$$

The strong approximations are fully characterized by P_1, P_2 and P_3 because:

$$\begin{aligned}
\underline{apr'}_{\mathbb{C}}(A) &= \cup \{K \in \mathbb{C} \mid (\forall X \in P_{\mathbb{C}})(X \subseteq K \Rightarrow X \in P_1(A))\} \\
\overline{apr'}_{\mathbb{C}}(A) &= \cup \{X \in P_{\mathbb{C}} \mid (\forall K \in \mathbb{C})(X \cap K \neq \emptyset \\
&\qquad \Rightarrow (\exists X' \in P_1(A) \cup P_2(A))(X' \subseteq K))\}
\end{aligned}$$

For a given tripartition (P_1, P_2, P_3) of $P_{\mathbb{C}}$ the number of sets $A \subseteq U$ that have $P_i(A) = P_i$ (for $i \in \{1, 2, 3\}$) is equal to:

$$\prod_{X \in P_2} \left(2^{|X|} - 2\right)$$

Because of this we can count all $A \in \mathcal{A}_k$ by generating all possible assignments (P_1, P_2, P_3) that have the correct lower and upper approximations.

For each $A \in \mathcal{A}_k$, every block $X \in P_{\mathbb{C}}$ is in some $P_i(A)$; we call i the type of X with respect to A. We will count all sets $A \in \mathcal{A}_k$, by deciding for each block $X \in P_{\mathbb{C}}$ which type it has. This needs to be done in such a way that every $Y \in \mathbb{C}$ is 'satisfied', meaning:

- If $Y \subseteq \underline{A}_k$, all blocks $X \subseteq Y$ must be in $P_1(A)$
- If $Y \cap \overline{A}_k = \emptyset$, all blocks $X \subseteq Y$ must be in $P_3(A)$
- Else, at least one block $X \subseteq Y$ should be in $P_1(A) \cup P_2(A)$ and at least one block $X \subseteq Y$ should be in $P_2(A) \cup P_3(A)$.

It is clear that every A for which these conditions hold has $(\underline{A}_k, \overline{A}_k)$ as its approximations.

Based on the first two cases above, for some blocks X we can immediately establish their type. For others (blocks that do not appear in the first two cases) we need to perform a backtracking search. Note that for every unsatisfied $Y \in \mathbb{C}$, $Y \cap \overline{A}_k \neq \emptyset$ and $Y \not\subseteq \underline{A}_k$.

The process proceeds as follows: as long as there are unsatisfied $Y \in \mathbb{C}$, we recursively pick one such Y arbitrarily and decide how it can be satisfied. For this, we arbitrarily order the blocks $X \subseteq Y$ which are still undecided. Then we vary the type of the first X among all types that are consistent with the requested lower and upper approximation. We repeat this process until Y is satisfied. Notice that this implies we can leave some $X \subseteq Y$ undecided. We also take into consideration that a block with only one element can never be in $P_2(A)$. Once all $Y \in \mathbb{C}$ are satisfied, we count the number of sets $A \subseteq U$ for which the type of each $X \in P_{\mathbb{C}}$ coincides with the assignment and add this to a running counter. That is:

$$\Pi_{X \in P_2}(2^{|X|} - 2) \cdot \Pi_{X \in N}(2^{|X|})$$

where N denotes the set of blocks for which no decision was made. By using recursion, we exhaustively generate all

possible assignments that satisfy every Y in $\mathbb{C}$ and thus we count all sets A with the right approximations.

The pseudo-code of this approach can be found in Algorithm 3. Below, we detail some optimizations that can be performed:

- When deciding the type of a block X for a $Y \in \mathbb{C}$ that already has some elements in P_1, it is enough to decide whether X is in P_1 or in $P_2 \cup P_3$, instead of deciding if it is in P_1, P_2 or P_3. Therefore the branching factor of the recursion is reduced from 3 to 2 in such cases.
- Instead of picking an arbitrary unsatisfied $Y \in \mathbb{C}$, we can choose a more specific Y in order to speed up the algorithm. A common heuristic is to make the next decision about the object with the fewest possibilities. In this case, we choose the Y with the least amount of undecided blocks left.
- The partition of U into blocks can be redefined dynamically during the runtime of the algorithm. Particularly if a $Y \in \mathbb{C}$ gets satisfied but has some undecided blocks left, it is no longer relevant that they are in Y, and therefore they could possibly be merged with other blocks. We chose not to implement this in general because it requires a lot of 'bookkeeping', but we used a limited variant. After the initial assignment of types to blocks that can be made without guessing, we make the partition coarser just once.

We will call the estimator of entropy that uses this approach to determine $|\mathcal{A}_k|$ the Horvitz-Thompson-exact estimator (HTEE).

Example 4. *Let U be a universe and $\mathbb{C}$ be the covering with three elements, shown in Figure 1, such that $P_{\mathbb{C}}$ has 5 blocks: $X_1, \ldots, X_5$ and $\mathbb{C} = \{X_1 \cup X_2, X_2 \cup X_3 \cup X_4, X_4 \cup X_5\}$.*

We want to calculate the number of sets $A \subseteq U$ so that $\underline{apr'}_{\mathbb{C}}(A) = X_4 \cup X_5$ and $\overline{apr'}_{\mathbb{C}}(A) = U$. We do this by exhaustively generating all possible assignments[1] of P_1, P_2, P_3 and N, using Algorithm 3 . These are shown in Table I. For example, for the first assignment, we see that even though the type of X_3 is undecided, all elements in $\mathbb{C}$ are satisfied. The amount of sets that are realizations of assignment 5 are:

$$\texttt{count}(\textit{assignment 5}) = \left(2^{|X_1|} - 2\right)\left(2^{|X_2|} - 2\right)2^{|X_3|}$$

If we add up the amount of realizations of all possible assignments we get the following result:

$$2^{|X_3|}\left(2^{(|X_1|+|X_2|)} - 2\right) - 2^{|X_1|} + 1$$

If we set the size of all blocks equal to 10 we have that the probability of a random subset of U having $(X_4 \cup X_5, U)$ as approximations is equal to:

$$\frac{2^{|X_3|}\left(2^{(|X_1|+|X_2|)} - 2\right) + 1}{2^{50}} = \frac{1073738753}{1125899906842624} \approx 9.537 \times 10^{-7}$$

[1]We assume that all blocks have at least two elements such that the assignment of type P_2 makes sense for each block.

Algorithm 3: Exact approximation counting

Input: covering $\mathbb{C}$; approximation $(\underline{A}, \overline{A})$
`// Preprocessing`
Check which $Y \in \mathbb{C}$ lie in $\underline{A}$, and which are disjoint with $\overline{A}$
Assign their elements the required type
$\mathbb{C}' \leftarrow \{Y \in \mathbb{C} \mid Y \text{ is still unsatisfied}\}$
Identify the 'blocks' with respect to $\mathbb{C}'$
$total \leftarrow 0$
`assign()`
Result: $total$ contains the number of sets with the given approximation

Function `assign()`**:**
  if *all $Y \in \mathbb{C}'$ are satisfied* **then**
    $total \leftarrow total + \Pi_{X \in P_2}(2^{|X|} - 2) \cdot \Pi_{X \in N}(2^{|X|})$
  else
    $Y \leftarrow$ unsatisfied $Y \in \mathbb{C}'$ with fewest amount of unassigned blocks
    `satisfy` (Y)
  end

Function `satisfy(`Y`)`**:**
  if Y *is satisfied* **then**
    `assign()`
  else
    $X \leftarrow$ first unassigned block in Y
    for $j \in \{1, 2, 3\}$ **do**
      $type_X \leftarrow P_j$
      `satisfy` (Y)
    end
    $type_X \leftarrow N$
  end

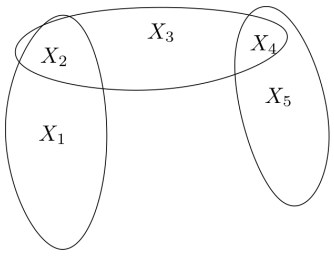

Fig. 1. $\mathbb{C}$ and its blocks from Example 4

Example 5. *We take Example 3 and inspect how the HTEE estimator will evaluate the entropy of $\mathbb{C}$. If we generate k sets of the universe U, we can safely assume that no collisions occur and no set is in the class of size two if k is small enough, since it is statistically very unlikely. Using the method described above, we can calculate the exact size of the equivalence class of each approximation that is generated*

TABLE I
ALL ASSIGNMENTS OF $P_{\mathbb{C}}$ GENERATED BY ALGORITHM 3 WHERE $\underline{apr'}_{\mathbb{C}}(A) = X_4 \cup X_5$ AND $\overline{apr'}_{\mathbb{C}}(A) = U$.

Assignments	1	2	3	4	5	6	7	8	9
P_1	X_1, X_4, X_5	X_1, X_4, X_5	X_2, X_4, X_5	X_2, X_4, X_5	X_4, X_5	X_4, X_5	X_2, X_4, X_5	X_2, X_4, X_5	X_4, X_5
P_2	X_2	$\emptyset$	X_1, X_3	X_1	X_1, X_2	X_1	X_3	$\emptyset$	X_2
P_3	$\emptyset$	X_2	$\emptyset$	X_3	$\emptyset$	X_2	X_1	X_1, X_3	X_1
N	X_3	X_3	$\emptyset$	$\emptyset$	X_3	X_3	$\emptyset$	$\emptyset$	X_3

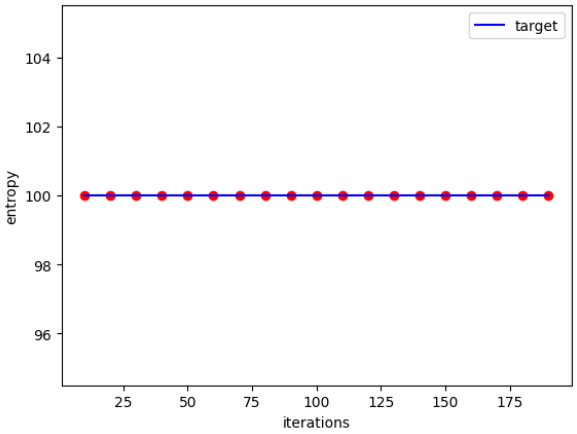

Fig. 2. The result of HTEE for the covering $\mathbb{C}$ with $n = 100$

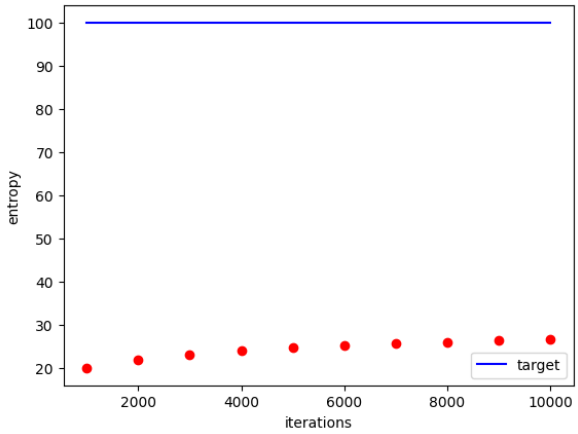

Fig. 3. The result of the coverage-adjusted estimator for the covering $\mathbb{C}$ with $n = 100$

and we get the following estimate:

$$\begin{aligned}\hat{\mathcal{H}} &= -\sum_{i=1}^{k} \frac{\log(2^{-n})}{2^n(1-(1-2^{-n})^k)} \\ &= \frac{kn}{2^n(1-(1-2^{-n})^k)} \\ &\approx \frac{kn}{2^n k 2^{-n}} \\ &= n\end{aligned}$$

The third line follows from:

$$\left(1-\frac{1}{2^n}\right)^k = \sum_{i=0}^{k}\binom{k}{i}(-2)^{ni} \approx 1-\frac{k}{2^n},$$

which holds when k is significantly smaller than 2^n.

V. EXPERIMENTS

In this section we will compare the previously discussed estimations.

We first evaluate the entropy for $n = 100$ for the covering from Example 3. When we use HTEE we see that, even after generating only 2 sets we already get an estimate that is less than 10^{-30} removed from $100 - \frac{1}{2^{-99}}$, as can be seen in Figure 2. The variance estimated by Equation (6) is practically zero. This confirms our findings from Example 5.

We may also observe that the coverage-adjusted estimator runs into problems because the value of the entropy is too large. Because of this there will be no collisions, which will cause $\hat{C}$ to be too far off. For completeness we provide a plot of the results of the coverage-adjusted estimator in Figure 3.

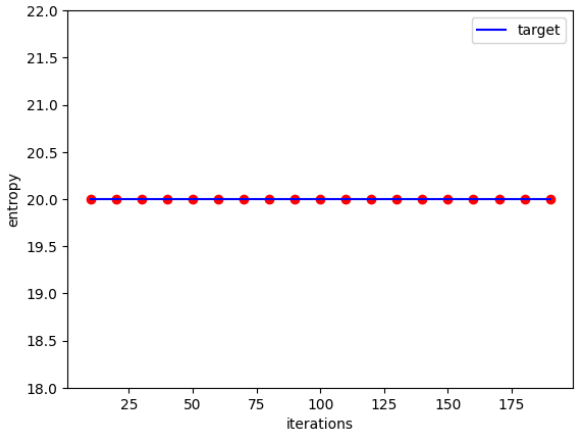

Fig. 4. The result of HTEE for the covering $\mathbb{C}$ with $n = 20$

When n is reduced to 20, both estimators produce accurate results, as seen in Figures 4 and 5.

Next, we consider real-world datasets from the UCI repository [33] and use coverings generated by Mapper. We use the following parameters for Mapper: the number of cubes is equal to 4, the overlap percentage is equal to 0.2, we use the projection onto the first PCA component as the lens function, and as a clustering algorithm we use the single-linkage algorithm with distance cutoff ε. The parameter ε will depend on the dataset we use since the average distance between points varies between datasets. We use the datasets with corresponding ε as displayed in Table II. The size of the

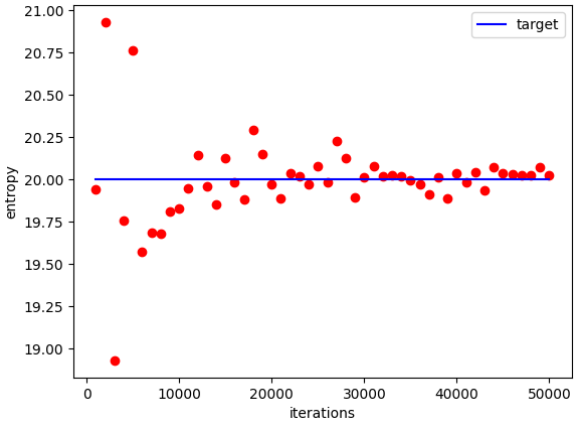

Fig. 5. The result of the coverage-adjusted estimator for the covering $\mathbb{C}$ with $n = 20$

TABLE II
THE DATASETS USED IN SECTION V

	Attributes	Instances	ε	$\lvert\mathbb{C}\rvert$
Iris	5	150	0.6	25
Vertebral Column	6	310	20	43
Wine	14	178	17.5	64

resulting coverings can also be found in this table.

The results obtained from the Iris dataset are presented in Figures 6 and 7. These figures show that the HTEE estimator converges rapidly after a small number of iterations. Additionally, the coverage-adjusted estimator consistently yields values lower than the actual entropy, which can be reasonably approximated as 15 based on the variance estimate of the HTEE estimator.

For the Vertebral Column and Wine datasets the entropy is again too high, meaning that we can only consider the HTEE estimator.

We see promising results where the variance of the estimator

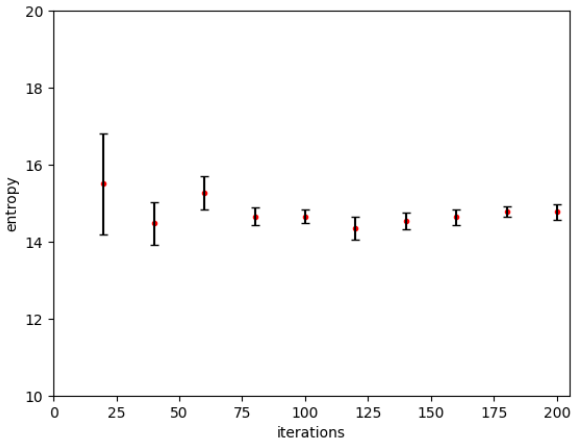

Fig. 6. The result of HTEE for the Iris dataset with $\varepsilon = 0.6$.

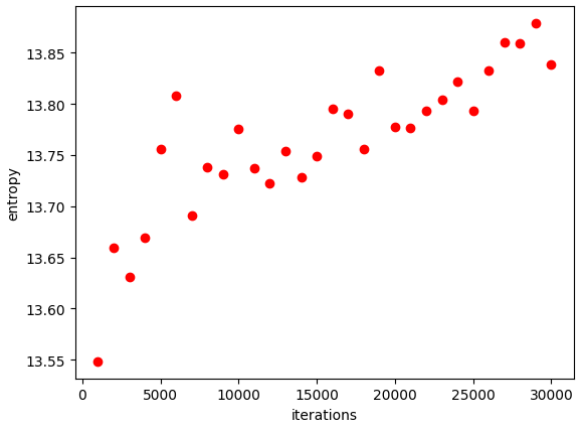

Fig. 7. The result of the coverage-adjusted estimator for the Iris dataset with $\varepsilon = 0.6$.

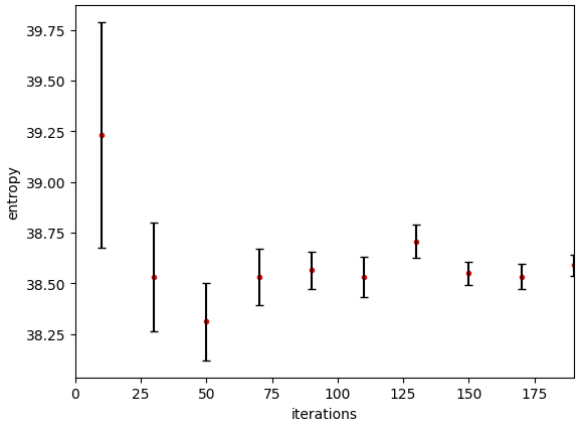

Fig. 8. The result of HTEE for the Vertebral Column dataset

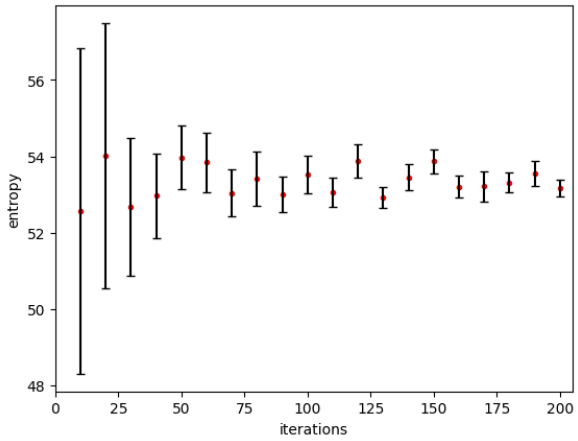

Fig. 9. The result of HTEE for the Wine dataset

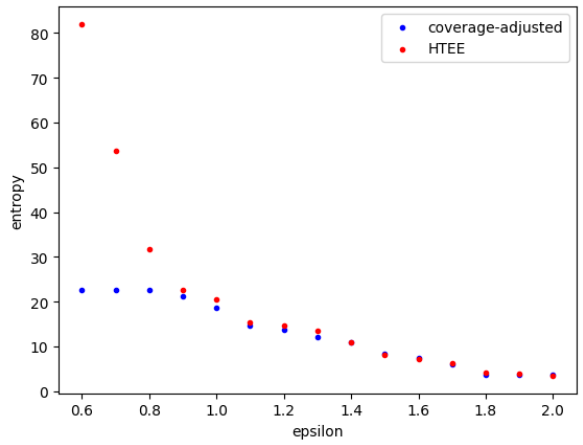

Fig. 10. The values of the HTEE and coverage-adjusted estimators for the iris dataset with a variable ε value.

quickly approaches zero as the number of iterations increases as shown in Figures 8 and 9.

We further analyze the Iris dataset by varying the value of ε, which controls the scale parameter in the Mapper algorithm. Adjusting ε generates a sequence of coverings under $\ll$, as mentioned in proposition 2. We vary ε from 0.6 to 2.0 in increments of 0.1. For the HTEE estimator, we perform 100 iterations, while for the coverage-adjusted estimator, we use 20,000 iterations—chosen to ensure comparable execution times for both methods. The variance of the HTEE estimator remains below 0.30 across all experiments. Based on this setup, Figure 10 presents the resulting comparisons. As shown, both estimators yield similar results when ε is large and the entropy is low. However, as entropy increases, the coverage-adjusted estimator consistently underestimates the true entropy. In cases of very high entropy, its estimates deviate significantly from the expected values.

VI. Discussion and Future Work

In this paper, we proposed an efficient method for estimating the entropy of the strong granule-based approximation operator in covering-based rough sets. Our approach involves generating random subsets of the universe U, determining the number of subsets that share the same rough set approximations, and applying the Horvitz-Thompson estimator to obtain an unbiased entropy estimate. We evaluated the performance of our estimator against the coverage-adjusted entropy estimator from the literature and found that our method not only yields greater accuracy but also performs more effectively in scenarios involving high-entropy coverings.

In future work, we plan to investigate estimates to calculate the size of $\mathcal{A}_k$ using properties of the covering and heuristics. A possible start to do this can be found in the following equation:

$$\Pi_{K\in P_{\mathbb{C}}, K\subseteq \overline{A}_k, K\not\subseteq \underline{A}_k}(2^{|K|}-2) \leq |\mathcal{A}_k| \leq 2^{|\overline{A}_k|-|\underline{A}_k|}$$

which can be proved by observing that

$$P_1 = \{K \in P_{\mathbb{C}} \mid K \subseteq \underline{A}_k\}$$
$$P_2 = \{K \in P_{\mathbb{C}} \mid K \subseteq \overline{A}_k, K \not\subseteq \underline{A}_k\}$$
$$P_3 = \{K \in P_{\mathbb{C}} \mid K \not\subseteq \overline{A}_k\}$$

may be used to assign types to all blocks.

When $\mathbb{C}$ is a partition, $\Pi_{K\in P_{\mathbb{C}}, K\subseteq \overline{A}_k, K\not\subseteq \underline{A}_k}(2^{|K|}-2)$ is exactly equal to $|\mathcal{A}_k|$. For general coverings, this does not hold any longer but we can use this as our approximation for $|\mathcal{A}_k|$.

Whenever $|K|$ is large enough for all K in $P_{\mathbb{C}}$, the difference between $2^{|K|}$ and $2^{|K|}-2$ is negligible. However in practice, most blocks in the used coverings are not that large. Because of this, we did not include this approach in the experimental analysis Section V. We still mention this estimate as it can probably be improved by using more heuristics to an even better estimate of $|\mathcal{A}_k|$.

Further research should also focus on constructing methods to estimate the entropy of (covering-based) rough set approximations other than the strong approximations. Finally, researchers and practitioners may also use these results to construct coverings with a given entropy. This is in particular useful for applications that require a certain degree of granularity.

Acknowledgments

This work was partially supported by Universidad Militar Nueva Granada, under project CIAS 3144-2020.

References

[1] Z. Pawlak, "Rough sets," *International Journal of Computer and Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982. doi: https://doi.org/10.1007/BF01001956

[2] Y. Yao, "On generalizing rough set theory," in *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 44–51.

[3] Z. Shi and Z. Gong, "The further investigation of covering-based rough sets: Uncertainty characterization, similarity measure and generalized models," *Information Sciences*, vol. 180, no. 19, pp. 3745–3763, 2010. doi: https://doi.org/10.1016/j.ins.2010.06.020

[4] T. Yang and Q. Li, "Reduction about approximation spaces of covering generalized rough sets," *International Journal of Approximate Reasoning*, vol. 51, no. 3, pp. 335–345, 2010. doi: https://doi.org/10.1016/j.ijar.2009.11.001

[5] I. Couso and D. Dubois, "Rough sets, coverings and incomplete information," *Fundam. Inform.*, vol. 108, pp. 223–247, 01 2011. doi: 10.3233/FI-2011-421

[6] J. Järvinen and S. Radeleczki, "Rough sets determined by tolerances," *International Journal of Approximate Reasoning*, vol. 55, no. 6, pp. 1419–1438, 2014. doi: https://doi.org/10.1016/j.ijar.2013.12.005

[7] D. Chen, X. Zhang, and W. Li, "On measurements of covering rough sets based on granules and evidence

theory," *Information Sciences*, vol. 317, pp. 329–348, 2015. doi: https://doi.org/10.1016/j.ins.2015.04.051

[8] L. D'eer and C. Cornelis, "Notes on covering-based rough sets from topological point of view: Relationships with general framework of dual approximation operators," *International Journal of Approximate Reasoning*, vol. 88, pp. 295–305, 2017. doi: https://doi.org/10.1016/j.ijar.2017.06.006

[9] M. Restrepo, C. Cornelis, and J. Gómez, "Partial order relation for approximation operators in covering based rough sets," *Information Sciences*, vol. 284, pp. 44–59, 2014. doi: https://doi.org/10.1016/j.ins.2014.06.032

[10] ——, "Duality, conjugacy and adjointness of approximation operators in covering-based rough sets," *International Journal of Approximate Reasoning*, vol. 55, no. 1, Part 4, pp. 469–485, 2014. doi: https://doi.org/10.1016/j.ijar.2013.08.002 Rough Sets and Logic.

[11] C. Wang, M. Shao, B. Sun, and Q. Hu, "An improved attribute reduction scheme with covering based rough sets," *Applied Soft Computing*, vol. 26, pp. 235–243, 2015. doi: https://doi.org/10.1016/j.asoc.2014.10.006

[12] L. Fachao, R. Yexing, and J. Chenxia, "Attribute reduction method of covering rough set based on dependence degree," *International Journal of Computational Intelligence Systems*, vol. 14, pp. 1419–1425, 2021. doi: 10.2991/ijcis.d.210419.002

[13] Y. Yang, D. Chen, X. Zhang, and Z. Ji, "Covering rough set-based incremental feature selection for mixed decision system," *Soft Comput.*, vol. 26, no. 6, pp. 2651–2669, 2022. doi: https://doi.org/10.1007/s00500-021-06687-0

[14] P. Zhu and Q. Wen, "Entropy and co-entropy of a covering approximation space," *International Journal of Approximate Reasoning*, vol. 53, no. 4, pp. 528–540, 2012. doi: https://doi.org/10.1016/j.ijar.2011.12.004

[15] Z. Li, S. Wei, and S. Liu, "Outlier detection using conditional information entropy and rough set theory," *Journal of Intelligent & Fuzzy Systems*, vol. 46, no. 1, pp. 1899–1918, 2024. doi: 10.3233/JIFS-236009

[16] L. Zhou and F. Jiang, "A rough set approach to feature selection based on relative decision entropy," in *Rough Sets and Knowledge Technology*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. doi: doi.org/10.1007/978-3-642-24425-4_17 pp. 110–119.

[17] M. Restrepo and C. Cornelis, "Mapper-based rough sets," in *Rough Sets: International Joint Conference, IJCRS 2024, Halifax, NS, Canada, May 17–20, 2024, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2024. doi: https://doi.org/10.1007/978-3-031-65665-1_1 p. 3–17.

[18] A. D. Smith, P. Dłotko, and V. M. Zavala, "Topological data analysis: Concepts, computation, and applications in chemical engineering," *Computers & Chemical Engineering*, vol. 146, 2021. doi: https://doi.org/10.1016/j.compchemeng.2020.107202

[19] V. Q. Vu, B. Yu, and R. E. Kass, "Coverage-adjusted entropy estimation," *Statistics in Medicine*, vol. 26, no. 21, pp. 4039–4060, 2007. doi: https://doi.org/10.1002/sim.2942

[20] J. G. Bazan, H. S. Nguyen, S. H. Nguyen, P. Synak, and J. Wróblewski, *Rough Set Algorithms in Classification Problem*. Heidelberg: Physica-Verlag HD, 2000, pp. 49–88.

[21] W. Ziarko and N. Shan, "Discovering attribute relationships, dependencies and rules by using rough sets," in *Proceedings of the Twenty-Eighth Annual Hawaii International Conference on System Sciences*, vol. 3, 1995. doi: 10.1109/HICSS.1995.375608 pp. 293–299 vol.3.

[22] H. Alqaheri, A. E. Hassanien, and A. Abraham, "Discovering stock price prediction rules using rough sets," *Neural Network World*, vol. 18, 01 2008.

[23] M. S. Raza and U. Qamar, *Understanding and using rough set based feature selection: Concepts, techniques and applications*, 2nd ed. Singapore, Singapore: Springer, Sep. 2019.

[24] K. Thangavel and A. Pethalakshmi, "Dimensionality reduction based on rough set theory: A review," *Applied Soft Computing*, vol. 9, no. 1, pp. 1–12, 2009. doi: https://doi.org/10.1016/j.asoc.2008.05.006

[25] A. K. Das, S. Sengupta, and S. Bhattacharyya, "A group incremental feature selection for classification using rough set theory based genetic algorithm," *Applied Soft Computing*, vol. 65, pp. 400–411, 2018. doi: https://doi.org/10.1016/j.asoc.2018.01.040

[26] W. De Maeyer, C. Cornelis, and M. Restrepo, "The granular structure of covering-based rough sets generated by mapper," in *Proceedings of 16th European Symposium on Computational Intelligence and Mathematics (ESCIM 2025)*, in press.

[27] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*, 5th ed., ser. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley-Blackwell, 2011.

[28] C. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[29] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, pp. 1191–1253, 2003. doi: https://doi.org/10.1162/089976603321780272

[30] A. Chao and T.-J. Shen, "Nonparametric estimation of shannon's diversity index when there are unseen species in sample. environ ecol stat 10: 429-443," *Environmental and Ecological Statistics*, vol. 10, pp. 429–443, 2003. doi: https://doi.org/10.1023/A:1026096204727

[31] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 663–685, 1952. doi: https://doi.org/10.2307/2280784

[32] M. Skorski, "Improved estimation of collision entropy in high and low-entropy regimes and applications to

anomaly detection," *IACR Cryptol. ePrint Arch.*, vol. 2016, p. 1035, 2016.

[33] M. Kelly, R. Longjohn, and K. Nottingham, the UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu