ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9415005

CONSTANT APPROXIMATION ALGORITHM FOR MINIMIZING CONCAVE IMPURITY

Thuan Nguyen, Hoang Le, and Thinh Nguyen

School of EECS, Oregon State University, Corvallis, OR 97331-5501, USA
nguyeth9@oregonstate.edu, lehoang@oregonstate.edu, thinhq@eecs.oregonstate.edu

ABSTRACT

Partitioning algorithms play a key role in many scientific and engineering disciplines. A partitioning algorithm divides a set into a number of disjoint subsets or partitions. Often, the quality of the resulted partitions is measured by the amount of impurity in each partition, the smaller impurity the higher quality of the partitions. Let M be the number of N-dimensional elements in a set and K be the number of desired partitions, then an exhaustive search over all the possible partitions to find a minimum partition has the complexity of $O(K^M)$ which quickly becomes impractical for many applications with modest values of K and M. Thus, many approximate algorithms with polynomial time complexity have been proposed, but few provide the bounded guarantee. In this paper, we propose a linear time algorithm with bounded guarantee based on the maximum likelihood principle. Furthermore, the guarantee bound of the proposed algorithm is better than the state-of-the-art method in [1] for many impurity functions, and at the same time, for $K \geq N$, the computational complexity is reduced from $O(M^3)$ to $O(M)$.

Index Terms— Partition, approximation, impurity.

1. INTRODUCTION

Partitioning plays a key role in many scientific and engineering disciplines. A partitioning algorithm divides a set of M N-dimensional elements into K disjoint subsets or partitions to optimize an objective function. Often, the quality of the resulted partitions is measured by the amount of impurity in each partition, the smaller impurity the higher quality of the partitions. Typically, the amount of impurity is measured by a real-valued function over the resulted partitions. In general, for a given impurity measure specified by a function over the partitions, finding the minimum impurity partitions is an NP-hard problem [2, 3]. Since the number of possible partitions is K^M, an exhaustive search over all the possible partitions to find a minimum partition has the complexity of $O(K^M)$ which quickly becomes impractical for many applications with modest values of K and M. To that end, many approximate algorithms with polynomial time complexity have been proposed, but few provide bounded guarantee [4–17]. Many of these algorithms exploit the concavity of the impurity function to speed up the running time [5], [8], [9], [17], [18]. For example, Burshtein et al. [8] and Coppersmith et al. [5] provided algorithms and theoretical analysis for the partitioning problem for a general concave impurity function called "frequency-weighted impurity". These "frequency-weighted impurity" are concave functions over its second argument. Two popular impurity functions the Gini index [9] and Shannon entropy [4] belong to this class of frequency-weighted impurity. Burshtein et al. and Coppersmith et al. showed that an optimal frequency-weighted impurity partition is separated by hyperplane cuts in the space of probability distributions. From this insight, they also proposed polynomial time algorithms to determine the optimal partitions [5], [8]. Based on the work of Burshtein et al., Kurkoski and Yagi proposed an algorithm to find the globally optimal partition that minimizes entropy impurity in $O(M^3)$ when $N = 2$ [19]. Although many heuristic algorithms have been proposed, there are few results in finding algorithms that provides a bounded guarantee on the performance. To fill this gap, recently Laber et al. [20] constructed a 2-approximation algorithm with the computational complexity of $O(2^N M \log M)$ for binary partition ($K = 2$). In other words, Laber et al. showed that the impurity achieved by their algorithm is at most a factor of 2 away from the true optimal impurity. As the extension of the work in [20], Cicalese et al. [1] proposed a heuristic algorithm for the number of partitions $K > 2$ based on dynamic programming technique in [19]. Their proposed algorithm runs in polynomial time $O(M^3)$ and can achieve $\log^2(\min\{N, K\})$-approximation for the entropy impurity and 3-approximation for the Gini index impurity.

In this paper, we propose a linear time algorithm with bounded guarantee based on the maximum likelihood principle for a wide class of impurity functions. To keep the generality of the impurity functions, instead of the providing a constant factor approximation, the proposed algorithm provides both the upper bound and the lower bound differently for different impurity functions. We show that for many well-known impurity functions such as entropy and Gini index, these bounds are theoretically better than that of the state-of-art method in [1] while the computational complexity is reduced from the polynomial time complexity $O(M^3)$ to the linear time complexity $O(M)$ for $K \geq N$.

2. PROBLEM FORMULATION

We assume that the data set $\mathbb{Y}$ to be partitioned consists of M discrete data points generated from an underlying probabilistic model. Specifically, let X be a discrete random variable taking on the values $x_1, x_2, \ldots, x_N$ with a given probability mass vector $\mathbf{p_x} = (p(x_1), p(x_2), \ldots, p(x_N))$. Let Y be another discrete random variable taking on values $y_1, y_2, \ldots, y_M$ which follows a given conditional probability $p(y_j|x_i)$. The goal is to design a mapping Q that partitions $\mathbb{Y}$ into K partitions $z_1, z_2, \ldots, z_K \in Z$ such that a given impurity function over the resulted partitions is minimized.

$$Q(Y) \rightarrow Z.$$

In this setting, for given $p(x_i)$ and $p(y_j|x_i)$, $p(x_i, y_j)$ are assumed to be given $\forall i, j$. Thus, each data point y_j is represented by a joint distribution vector $\mathbf{p}_{\mathbf{x},y_j} = (p(x_1, y_j), p(x_2, y_j), \ldots, p(x_N, y_j))$. Each mapping Q induces a joint distribution vector $\mathbf{p}_{\mathbf{x},z_k} = (p(x_1, z_k), p(x_2, z_k), \ldots, p(x_N, z_k))$ between X and $Z = z_k$. The conditional distribution of X given Z $p(x_i|z_k)$ and the marginal probability mass function of Z $p(z_k)$ can be determined from $\mathbf{p}_{\mathbf{x},z_k}$. We want to find an optimal mapping Q^* to minimize the impurity function I_Q that satisfies two following conditions:

- (Required) I_Q has the following form:

$$I_Q = \sum_{k=1}^{K} \sum_{i=1}^{N} p(z_k) f(p(x_i|z_k)), \tag{1}$$

  where $f(.) : \mathbb{R} \rightarrow \mathbb{R}^+$ is a non-negative concave function.

- (Optional) $f(x) = xl(x)$ where $l(x) : \mathbb{R} \rightarrow \mathbb{R}$ is a convex function. This second condition is optional in the sense that we use it in the analysis of the constant factor approximation for the proposed algorithm. The algorithm itself does not make use of this condition.

Many popular impurity functions such as entropy and Gini index satisfy our conditions. Noting that in [1] and [20], to guarantee the constant factor approximation, the authors considered a class of impurity concave functions $f(.)$ with an additional condition on $xf''(x)$ being a non-increasing function.

3. IMPURITY MINIMIZATION ALGORITHM

In this section, we first construct both upper and lower bounds for impurity functions of the form in (1). Using these bounds, we show that the proposed maximum likelihood algorithm achieves a constant factor approximation. First, define:

$$k^* = \underset{1 \leq i \leq N}{\operatorname{argmax}} \, p(x_i|z_k), \tag{2}$$

$$e_Q = \sum_{k=1}^{K} p(z_k) p(x_{k^*}|z_k), \tag{3}$$

and

$$e^{\max} = \max_{Q} e_Q. \tag{4}$$

For a given k, x_{k^*} is most likely to produce z_k. Therefore, e_Q is the weighted sum of the maximum likelihood of each x_{k^*} for each z_k. In addition, from (2), $1/N \leq e_Q \leq 1$. We also note that each mapping Q induces a $p(x_i, z_k)$ and thus $p(x_i|z_k)$. So k^* and e_Q are different for different Q. Our approach to find the minimum impurity is to find two functions: $u(e_Q)$ and $l(e_Q)$ such that $l(e_Q) \leq I_Q \leq u(e_Q)$. Furthermore, we show that $u(e_Q)$ and $l(e_Q)$ are decreasing functions for many impurities. Therefore, by minimizing $u(e_Q)$, i.e., maximizing e_Q, we can bound the minimum value of I_Q between $u(e_Q)$ and $l(e_Q)$ for some e_Q.

3.1. Upper Bound of The Impurity Function

We have the following theorem for the upper bound of an impurity function I_Q.

Theorem 1. (**Upper bound**) For any given mapping Q that induces e_Q, let

$$u(e_Q) = f(e_Q) + (N-1) f(\frac{1 - e_Q}{N-1}), \tag{5}$$

then:

$$u(e_Q) \geq I_Q. \tag{6}$$

Proof. From the definition of the impurity function, we have:

$$\begin{aligned}
I_Q &= \sum_{k=1}^{K} \sum_{i=1}^{N} p(z_k) f(p(x_i|z_k)) \\
&= \sum_{k=1}^{K} p(z_k) f(p(x_{k^*}|z_k)) + \sum_{k=1}^{K} \sum_{i \neq k^*, i=1}^{N} p(z_k) f(p(x_i|z_k)) \\
&\leq f(\sum_{k=1}^{K} p(z_k) p(x_{k^*}|z_k)) + \sum_{k=1}^{K} \sum_{i \neq k^*, i=1}^{N} p(z_k) f(p(x_i|z_k)) \quad (7) \\
&\leq f(\sum_{k=1}^{K} p(z_k) p(x_{k^*}|z_k)) \\
&\quad + \sum_{k=1}^{K} p(z_k) [(N-1) f(\frac{\sum_{i=1, i \neq k^*} p(x_i|z_k)}{N-1})] \quad (8) \\
&= f(e_Q) + (N-1) \sum_{k=1}^{K} p(z_k) f(\frac{1 - p(x_{k^*}|z_k)}{N-1}) \quad (9) \\
&\leq f(e_Q) + (N-1) f(\frac{\sum_{k=1}^{K} p(z_k)(1 - p(x_{k^*}|z_k))}{N-1}) \quad (10) \\
&= f(e_Q) + (N-1) f(\frac{1 - e_Q}{N-1}), \quad (11)
\end{aligned}$$

where (7) is due to concavity of $f(.)$ and $\sum_{k=1}^{K} p(z_k) = 1$, (8) is due to Jensen inequality for concave function, (9) is due to the definition of e_Q and $\sum_{i=1, i \neq k^*}^{N} p(x_i|z_k) + p(x_{k^*}|z_k) = 1$, (10) is due to concavity of $f(.)$ together with $\sum_{k=1}^{K} p(z_k) = 1$, (11) is due to $\sum_{k=1}^{K} p(z_k) = 1$. If $f(.)$ is entropy function, our bound is identical to Fano's inequality [21]. □

Theorem 2. $u(e_Q)$ is a monotonic decreasing function.

Proof. By taking the derivative of $u(e_Q)$ and noting that $e_Q \geq 1/N, \forall Q$, it is possible to show that $u'(e_Q) < 0$ or $u(e_Q)$ is a monotonic decreasing function. □

Based on Theorem 2, let $e^{\max}$ be the maximum value over all e_Q i.e., $e^{\max} = \max_Q e_Q$, then $u(e^{\max})$ has the minimum value. Since $u(e_Q)$ is an upper bound of I_Q, $u(e^{\max})$ provides a good upper bound for I_{Q^*}. We now state an important result that characterizes the structure of the $e^{\max}$ mapping ($Q_{e^{\max}}$).

Theorem 3. (**Structure of the** $\mathbf{e}^{\max}$ **mapping**) Let $\mathcal{Z}$ and $\mathcal{X}$ be the sample spaces of Z and X, respectively. Let $j^* = \operatorname{argmax}_i p(x_i, y_j)$ and define mapping $Q_{e^{\max}}$ with the following structure:

$$Q_{e^{\max}}(y_j) = z_{j^*}. \tag{12}$$

(a) If $|\mathcal{Z}| = |\mathcal{X}|$, then $Q_{e^{\max}}$ produces $e^{\max} = \max_Q e_Q$. Conversely, for any Q that produces $e^{\max}$, Q must have the structure of $Q_{e^{\max}}$.

(b) If $|\mathcal{Z}| > |\mathcal{X}|$, then $Q_{e^{\max}}$ still produces $e^{\max} = \max_Q e_Q$. However, it is not necessary that for any Q that produces $e^{\max}$, Q must have the structure of $Q_{e^{\max}}$.

Proof. Please see our extension version. □

3.2. Algorithm

Based on the upper bound in Theorem 1, to minimize the impurity function, one wants to minimize the impurity's upper bound $u(e_Q)$. Based on Theorem 2, to minimize $u(e_Q)$, one wants to maximize e_Q. Let $\mathcal{V}_K$ be the set of binary N-dimensional vectors $\mathbf{v}$'s, each contains exactly K entries 1 and $N - K$ entries 0. Thus, the size of $\mathcal{V}_K$ is $\binom{N}{K}$. For each $\mathbf{v} = (v_1, v_2, \ldots, v_N)$, define the N-dimensional vector: $\mathbf{p}'_{\mathbf{x},y_j} = (v_1 p(x_1, y_j), v_2 p(x_2, y_j), \ldots, v_N p(x_N, y_j))$ then $\mathbf{p}'_{\mathbf{x},y_j}$ has exactly K non-zero entries. Next, we consider the following possible cases.

1. $K = N$: When $K = N$, $\mathcal{V}_K = \mathcal{V}_N$ contains exactly one $\mathbf{v}$ which is $\mathbf{v} = (1, 1, \ldots, 1)$. Using Theorem 3-(a) with $p(x_i, y_j)$ replaced by $p'(x_i, y_j)$ will produce $e^{\max}$.
2. $K < N$: When $K < N$, there are $\binom{N}{K}$ mappings Q that partition K-dimension vectors $\mathbf{p}'_{\mathbf{x},y_j}$ to K partitions. Moreover, from the necessary condition in Theorem 3-(a), at least one of mapping in this $\binom{N}{K}$ mappings must achieve $e^{\max}$.
3. $K > N$: From Theorem 3-(b), the partition which achieves $e^{\max}$ is exactly the same with the partition when $K = N$.

Based on these possible cases and using Theorem 3, the algorithm follows.

Running time of Algorithm 1: To find the partition that generates $e^{\max}$, we need to search over all the possible mappings $\mathbf{v} \in \mathcal{V}_K$. For each $\mathbf{v}$, Algorithm 1 has complexity of $O(M)$. Since there are $\binom{N}{K}$ possible $\mathbf{v}$ if $K < N$, Algorithm 1 has the complexity of $O(\binom{N}{K} M)$. In the worst case when $K = N/2$, the complexity of Algorithm 1 is $O(2^{N/2} M)$. However, if $K \geq N$, there is only one mapping $\mathbf{v}$ and the running time of algorithm is $O(M)$.

4. CONSTANT FACTOR APPROXIMATION ANALYSIS FOR ENTROPY AND GINI INDEX

In this section, we state a few results for establishing the constant approximation property of Algorithm 1. The following theorem establishes a lower bound for I_Q. This lower bound predicates on the second condition $f(x) = x l(x)$ where $l(x)$ is a convex function. It is not used explicitly in the algorithm but is used in the analysis to establish the constant factor approximation property of the algorithm.

Algorithm 1 Finding $Q_{e^{\max}}$ and $e^{\max}$.

```
Input: Dataset $\mathbb{Y} = \{y_1, \ldots, y_M\}$, $p(x_i, y_j)$, $K$, and $N$.
Output: Partition $Z = \{z_1, z_2, \ldots, z_K\}$.
If $K < N$: $\mathcal{V} = \mathcal{V}_K$
If $K \geq N$: $\mathcal{V} = \mathcal{V}_N$
    For $\mathbf{v} \in \mathcal{V}$
        For $1 \leq j \leq M$, $1 \leq i \leq N$
            Step 1: Projection.
                $p'(x_i, y_j) = v_i p(x_i, y_j)$.        (13)
            Step 2: Finding the maximum likelihood.
                $j^* = \underset{1 \leq i \leq N}{\operatorname{argmax}} \{p'(x_i, y_j)\}$.        (14)
            Step 3: Partition assignment.
                $Q(y_j) \rightarrow z_{j^*}$.        (15)
        End For
        Computing $e^{\max}$: Using the resulted partitions
          $Z = \{z_1, z_2, \ldots, z_K\}$ and (3) to compute $e^{\max}$.
    End For
Return: Returning $Z = \{z_1, z_2, \ldots, z_K\}$ and $e^{\max}$.
```

Theorem 4. (**Lower bound**) For any given mapping Q that induces e_Q, we have:

$$I_Q \geq l(e_Q). \tag{16}$$

Proof. Please see our extension version. □

Theorem 5. ($\mathbf{R}(\mathbf{e}^{\max})$**-approximation**) Algorithm 1 provides $R(e^{\max})$-approximation for both entropy and Gini index impurities where:

$$R(e^{\max}) = \frac{u(e^{\max})}{l(e^{\max})}. \tag{17}$$

Proof. Let I_{Q^*} be the minimum impurity and $I_{Q_{e^{\max}}}$ be the impurity produced by running Algorithm 1. Now, assume that Q^* produces e_{Q^*}. From the definition of $e^{\max}$, $e_{Q^*} \leq e^{\max}$. Moreover, it is straightforward to show that $l(e_Q)$ for both entropy and Gini index impurities are decreasing functions. Thus, $I_{Q^*} \geq l(e_{Q^*}) \geq l(e^{\max})$. Therefore,

$$\frac{I_{Q_{e^{\max}}}}{I_{Q^*}} \leq \frac{u(e^{\max})}{\min_{e_Q} l(e_Q)} = \frac{u(e^{\max})}{l(e^{\max})} = R(e^{\max}). \tag{18}$$

Thus, the impurity produced by Algorithm 1 is guaranteed to be away from the true solution by at most a factor of $R(e^{\max})$. □

The result in Theorem 5 can be applied for any concave impurity function $f(x) = x l(x)$ with $l(x)$ being a non-increasing function. Next, we show that $R(e^{\max})$-approximation is better than the approximation in [1] for both the entropy impurity and the Gini index impurity.

Theorem 6. Algorithm 1 provides a 2-approximation for Gini index impurity.

Impurity	Data set	K	$\mathrm{e}^{\max}$	$\mathrm{R}(\mathrm{e}^{\max})$	Alg. 1	Iter-Alg.	Alg. 1/Iter-Alg.
Entropy	20NEWS	20	0.2420	1.9630	3.9658	3.9043	1.01575
	RCV1	103	0.2185	2.7215	4.7935	4.5667	1.04966
Gini index	20NEWS	20	0.2420	1.2021	0.9055	0.8890	1.01856
	RCV1	103	0.2185	1.2108	0.9206	0.9052	1.01701

Table 1: Entropy and Gini index impurities using 20NEWS and RCV1 data sets.

Proof. For the Gini index impurity function, $f(x) = x(1-x)$ and $l(x) = 1 - x$. Thus,

$$R(e^{\max}) = \frac{f(e^{\max}) + (N-1)f(\frac{1-e^{\max}}{N-1})}{l(e^{\max})}$$

$$= \frac{e^{\max}(1-e^{\max}) + (N-1)\frac{1-e^{\max}}{N-1}(1-\frac{1-e^{\max}}{N-1})}{1-e^{\max}} \quad (19)$$

$$= e^{\max} + 1 - \frac{1-e^{\max}}{N-1} \leq e^{\max} + 1 \leq 2, \quad (20)$$

with (19) due to $f(x) = x(1-x)$ and $l(x) = 1-x$, (20) due to a bit of algebra and $e^{\max} \leq 1$. Noting that one can use $e^{\max}+1$ as another approximation for Gini index impurity. □

Remark 1. Algorithm 1 provides a 2-approximation for Gini index impurity in comparison of a 3-approximation in [1].

Theorem 7. The entropy impurity approximation provided by Algorithm 1 is better than the approximation in [1] in case of $K \geq N$, i.e., $R(e^{\max}) < \log^2(\min\{N, K\}) = \log^2(N)$ if

$$N \geq N^{\min} = 2^{S(e^{\max})}, \quad (21)$$

where

$$S(e^{\max}) = \frac{1-e^{\max}}{-2\log(e^{\max})} + \frac{\sqrt{4H(e^{\max})(-\log(e^{\max})) + (1-e^{\max})^2}}{-2\log(e^{\max})},$$

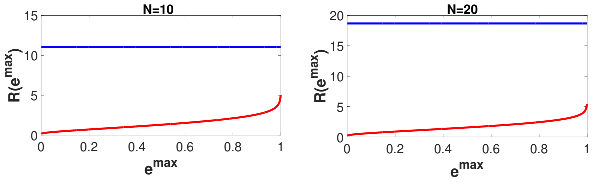
and $H(x) = -(x \log x + (1-x)\log(1-x))$ is the binary entropy of x.

Proof. Please see our extension version. □

If $K \geq N$, our bound for entropy impurity is theoretically better than that of [1] while the computational complexity is reduced from $O(M^3)$ to $O(M)$. When $K < N$, it is possible that the algorithm in [1] provides a better approximation. Fig. 1 shows the performance bound of the proposed algorithm vs. the state of the art in [1]. $R(e^{\max})$ vs. $e^{\max} \in (0.01, 0.99)$ for $N = K = 10$ and $N = K = 20$ are plotted in red while the approximations of [1] ($\log^2(N)$) are plotted in blue. As seen, the red curves are always below the blue curves. Moreover, the gaps between our approximation and that of [1] are proportional to the size of N. That said, for large values of N, our approximation is progressively better than that of [1]. We also note that $S(e^{\max})$ is monotonic increasing. Thus, if $e^{\max}$ increases, then $N^{\min}$ increases. For example, if $e^{\max} = 0.5$ then (21) holds for $N^{\min} = 2.42$, if $e^{\max} = 0.8$, (21) holds for $N^{\min} = 3.58$, if $e^{\max} = 0.9$, (21) holds for $N^{\min} = 4.34$, if $e^{\max} = 0.999$, (21) holds for $N^{\min} = 9.06$.

5. NUMERICAL RESULTS

To evaluate the performance of the proposed algorithm, we used two data sets: 20NEWS and RCV1 [22]. These are widely used for evaluating text classification methods. Existing algorithms [5], [10], [23] can only find locally optimal solutions. To approximate a globally optimal solution, many iterative algorithms use multiple random starting points and select the best solution. To that end, we compare the impurity provided by Algorithm 1 with the impurity produced by running the iterative algorithms from 100 randomly starting points. Although these iterative algorithms do not guarantee to find a globally optimal solution, their performances were shown in [23] to outperform the clustering methods in [24] and [25]. The code as well as the datasets are available at `https://github.com/hoangle96/linear_clustering`. Particularly, the dataset 20NEWS contains $M = 51840$ vectors of dimension $N = 20$ while the dataset RCV1 has 170946 vectors of dimension 103. The joint distribution $p(x_i, y_j)$ for these data set is computed ahead of time. We run both Algorithm 1 and the iterative algorithm using $K = 20$ for 20NEWS and $K = 103$ for RCV1. The impurity of these algorithms are provided in Table 1. As seen, the impurity provided by Algorithm 1 is very close to the impurity obtained from the iterative algorithm (see *Alg.1/Iter-Alg.* column in Table 1) (assuming that the iterative algorithm obtains a globally optimal solution). For the entropy impurity, the running times of Algorithm 1 for 20NEWS and RCV1 data sets are 0.02 and 0.03 seconds. These are significantly faster than the running times of iterative algorithm which are 83.57 and 1350.67 seconds, respectively. For the Gini index impurity, the running times of Algorithm 1 for 20NEWS and RCV1 data sets are 0.01 and 0.02 seconds, while the running times of iterative algorithm are 14.90 and 82.39 seconds, respectively. Again, Algorithm 1 is significantly faster.



(a) (b)

Fig. 1: $R(e^{\max})$ for entropy using (a) $N = 10$; (b) $N = 20$.

6. CONCLUSION

In this paper, we proposed a guaranteed bounded linear time algorithm for minimizing a wide class of impurity function including entropy and Gini index. In some cases, we showed that the proposed algorithm is better than the state-of-art algorithms in both terms of computational complexity and the quality of partitioned outputs.

7. REFERENCES

[1] Ferdinando Cicalese, Eduardo Laber, and Lucas Murtinho. New results on information theoretic clustering. In *International Conference on Machine Learning*, pages 1242–1251, 2019.

[2] Brendan Mumey and Tomáš Gedeon. Optimal mutual information quantization is np-complete. In *Proc. Neural Inf. Coding (NIC) Workshop*, 2003.

[3] F. Cicalese and E. S. Laber. Information theoretical clustering is hard to approximate. *IEEE Transactions on Information Theory*, pages 1–1, 2020.

[4] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[5] Don Coppersmith, Se June Hong, and Jonathan RM Hosking. Partitioning nominal attributes in decision trees. *Data Mining and Knowledge Discovery*, 3(2):197–217, 1999.

[6] Philip A. Chou. Optimal partitioning for classification and regression trees. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (4):340–354, 1991.

[7] Arthur Nádas, David Nahamoo, Michael A Picheny, and Jeffrey Powell. An iterative'flip-flop'approximation of the most informative split in the construction of decision trees. In *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, pages 565–568. IEEE, 1991.

[8] David Burshtein, Vincent Della Pietra, Dimitri Kanevsky, Arthur Nadas, et al. Minimum impurity partitions. *The Annals of Statistics*, 20(3):1637–1646, 1992.

[9] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[10] Jiuyang Alan Zhang and Brian M Kurkoski. Low-complexity quantization of discrete memoryless channels. In *2016 International Symposium on Information Theory and Its Applications (ISITA)*, pages 448–452. IEEE, 2016.

[11] T. Nguyen and T. Nguyen. Capacity achieving quantizer design for binary channels. *IEEE Communications Letters*, pages 1–1, 2020.

[12] Thuan Nguyen and Thinh Nguyen. On binary quantizer for maximizing mutual information. *IEEE Transactions on Communications*, pages 1–1, 2020.

[13] T. Nguyen and T. Nguyen. Communication-channel optimized impurity partition. In *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, pages 1–5, 2020.

[14] Thuan Nguyen and Thinh Nguyen. Optimal quantizer structure for binary discrete input continuous output channels under an arbitrary quantized-output constraint. *International Symposium on Information Theory (ISIT)*, 2020.

[15] T. Nguyen and T. Nguyen. Thresholding quantizer design for mutual information maximization under output constraint. In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pages 1–5, 2020.

[16] T. Nguyen and T. Nguyen. On thresholding quantizer design for mutual information maximization: Optimal structures and algorithms. In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pages 1–5, 2020.

[17] Thuan Nguyen and Thinh Nguyen. A linear time partitioning algorithm for frequency weighted impurity functions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5375–5379. IEEE, 2020.

[18] Thuan Nguyen and Thinh Nguyen. Minimizing impurity partition under constraints. *Transaction on Communications*, 2020, submitted.

[19] Brian M Kurkoski and Hideki Yagi. Quantization of binary-input discrete memoryless channels. *IEEE Transactions on Information Theory*, 60(8):4544–4552, 2014.

[20] Eduardo S Laber, Marco Molinaro, and Felipe A Mello Pereira. Binary partitions with approximate minimum impurity. In *International Conference on Machine Learning*, pages 2860–2868, 2018.

[21] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[22] Inderjit S Dhillon, Subramanyam Mallela, and Rahul Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of machine learning research*, 3(Mar):1265–1287, 2003.

[23] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

[24] Noam Slonim and Naftali Tishby. The power of word clusters for text classification. In *23rd European Colloquium on Information Retrieval Research*, volume 1, page 200, 2001.

[25] L Douglas Baker and Andrew Kachites McCallum. Distributional clustering of words for text classification. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 96–103, 1998.