Theoretical bounds on the error of M-ary LOM tree based on Gini and Modified Gini entropy

Weixi Zhang, wz1219, N18406096
Hongmiao Zhao, hz1509, N13066872

May 3, 2018

Abstract

We analyze the relation between multi-class classification error and Gini entropy and its modified version. Past work showed that under weak hypothesis assumption, we can limit the error of the decision tree to a bound by estimating the Shannon entropy after a certain number of splits. We proved that Gini entropy and its modified version can also be upper bounds of the error. These two bounds both can limit the error to a lower level. This result agrees with the numerical result in Choromanska, Choromanski & Bojarski, 2016. We also showed that why Gini entropy mimics the behavior of the error.

1. background

We commonly reduce extrem multi-class classification problem to a series of choices in tree-based modle, and each leaf of the tree represents a class. Because this kind of apporach allows for fast prediction.

In previous work, Choromanska and langford perposed a new tree molde, LOM tree algorithm (Choromanska&langford, 2015) and correspongding objective function. It achieves O(log(k)) computational time per example for both training and testing. Also, Choromanska and Choromanski explored the connection between objective function and three well-known entropy-based decision tree ovjectives, which is used to estimate the qulity of a tree. And they showed that maximaizing the considered objective function results also in the reduction of all thes entropy-based objectives. However the model above is a binary tree. Inspired by these two papers, Choromanska and Jernite present an objective function (Choromanska&Jernite, 2017) which favors high-quality node splits, combined balancedness with purity, for k-ary tree in order to reduce the depth of the tree. fururthermore, it proved that minimizing entropy-based objective, Shannon entropy, leads to reduction of overall classification error.

Our project report is an extension based on all materials above. We find the connection between multi-class classification error and entropy-based objectives, Gini-entropy and its modified version. We show how fast the multi-class classification error is reduced when we optimize classification error in each node of the classification tree. The main theoretical analysis of theis artical relies on the assumption of the existence of weak learners in the tree nodes as all previous papers.

This report is organized as follows: Section 2 outlines necessary algorithm and results in previous work, Section 3 contains our work and proof, section 4 comes to a conclusion.

2. previous related work

2.1 learning tree-strctured objectives

Choromanska defined the node objective J_n for node n in k-ary tree as (equation 6, Choromanska&Jernite, 2017) :

$$J_n = \frac{2}{M} \sum_{i=1}^{K} q_i^{(n)} \sum_{j=1}^{M} |p_j^{(n)} p_{j|i}^{(n)}| \tag{1}$$

where M is arity of the tree, $q_i^{(n)}$ denotes the proportion of nodes reaching node n that are of class i, $p_{j|i}^{(n)}$ is the probability that an example of class i reaching n will be sent to its jth child, and $p_j^{(n)}$ is the probability that an example of any class reaching n will be sent to its jth child. Note that we have:

$$\forall j \in [1, M], p_j^{(n)} = \sum_{i=1}^{K} p_j^{(n)} p_{j|i}^{(n)} \tag{2}$$

The objective in Equation 1 reduces to the LOM tree objective in the case of M = 2 (equation 1, Choromanska&langford, 2015).

2.1 entropy-based crieria

Three well-known entropy-based crierias are defined as:

Shannon entropy:

$$G_t^e = \sum_{l \in L} \omega_l \sum_{i=1}^{K} q_i^{(l)} \ln(\frac{1}{q_i^{(l)}}) \tag{3}$$

Gini-entropy:

$$G_t^g = \sum_{l \in L} \omega_l \sum_{i=1}^{K} q_i^{(l)} (1 - q_i^{(l)}) \tag{4}$$

Modified gini-entropy (where C is a constant such that $C > 2$:

$$G_t^g = \sum_{l \in L} \omega_l \sum_{i=1}^{K} \sqrt{q_i^{(l)}(C - q_i^{(l)})} \tag{5}$$

Theorem 1. (Theorem 1, Choromanska&Jernite, 2017) says that under the Weak Hypothesis Assumption (defination 5,Choromanska&langford, 2015), for any $k \in [0, 1]$, to obtain $\epsilon(T) < k$ ($\epsilon(T)$ is error of the tree T), it suffices to have a tree with $N >= (\frac{1}{k})^{\frac{16[M(1-2\gamma)+2\gamma](M-1)}{M^2\gamma^2} \ln K}$ internal nodes.
Where $\gamma \in [\frac{M}{2} \min_{j=1,2,\ldots,M} p_j, 1 - \frac{M}{2} \min_{j=1,2,\ldots,M} p_j], Jn \geq \gamma$

This theorem is only based on Shannon entropy. In other words. It only links multi-class classification error to Shannon entropy. In next section we will find the connection between multi-class classification error and other two entropy-based criterias and we will also as well prove it.

3. Theoretical Results and proof

Theorem 2. under the Weak Hypothesis Assumption, for any $k \in [0, 1]$, to obtain $\epsilon(T) < k$ ($\epsilon(T)$ is error of the tree T), it suffices to have a tree with $N >= (\frac{1}{k})^{\frac{8K[M(1-2\gamma)+2\gamma](1-\frac{1}{K})(M-1)}{M^2\gamma^2 \log_2 e}}$ internal nodes when we use Gini-entropy to estimate the quality of the tree (under the same condition as Theorem 1).

Proof

Notice that $q = \sum_M^{j=1} p_j q_j$. Then We denote the difference between the contribution of node n to the value of the entropy-based objectives in times t and $t + 1$ as:

$$\Delta_g^t = G_t^g - G_{t+1}^g = \omega[\tilde{G}^g(q) - \sum_{j=1}^{M} p_j \tilde{G}^g(q_j)] \tag{6}$$

Then we extend the inequality given by Equation 7 in (Choromanska et al., 2016) by applying Theorem5.2. from (Azocaretal.,2011) and obtain the following bound

$$\begin{aligned}
\Delta_g^t &= \omega[\tilde{G}^g(q) - \sum_{j=1}^{M} p_j \tilde{G}^g(q_j)] \\
&\geq \omega \sum_{j=1}^{M} p_j ||q_j - \sum_{l=1}^{M} p_l q_l||_2^2 \\
&\geq \frac{\omega}{K} \sum_{j=1}^{M} p_j ||q_j - \sum_{l=1}^{M} p_l q_l||_1^2 \\
&= \frac{\omega}{K} \sum_{j=1}^{M} p_j (\sum_{i=1}^{K} |\frac{q_i p_{j|i}}{p_j} - \sum_{l=1}^{M} p_l \frac{q_i p_{l|i}}{p_l}|)^2 \\
&= \frac{\omega}{K} \sum_{j=1}^{M} \frac{1}{p_j} (\sum_{i=1}^{K} |q_i p_{j|i} - p_j q_i \sum_{l=1}^{M} p_{l|i}|)^2 \\
&= \frac{\omega}{K} \sum_{j=1}^{M} \frac{1}{p_j} (\sum_{i=1}^{K} q_i |p_{j|i} - p_j|)^2
\end{aligned} \tag{7}$$

Notice that $p_j \leq \frac{M(1-2\gamma)+2\gamma}{M}$. So we get:

$$\begin{aligned}
\Delta_g^t &\geq \frac{M\omega}{K(M(1-2\gamma)+2\gamma)} \sum_{j=1}^{M} (\sum_{i=1}^{K} q_i |p_{j|i} - p_j|)^2 \\
&= \frac{M^2\omega}{K(M(1-2\gamma)+2\gamma)} \sum_{j=1}^{M} \frac{1}{M} (\sum_{i=1}^{K} q_i |p_{j|i} - p_j|)^2 \\
&\geq \frac{M^2\omega}{4K(M(1-2\gamma)+2\gamma)} (\sum_{j=1}^{M} \frac{2}{M} \sum_{i=1}^{K} q_i |p_{j|i} - p_j|)^2 \\
&= \frac{M^2\omega}{4K(M(1-2\gamma)+2\gamma)} Jn^2
\end{aligned} \tag{8}$$

where the last inequality is a consequence of Jensens inequality. ω can further be lower-bounded by noticing the following:

$$\begin{aligned}
G_t^g &= \sum_{l \in L} \omega_l \sum_{i=1}^{K} q_i^{(l)}(1 - q_i^{(l)}) \\
&\leq \omega \sum_{l \in L} \sum_{i=1}^{K} q_i^{(l)}(1 - q_i^{(l)}) \\
&= \omega \sum_{l \in L}(1 - \sum_{i=1}^{K}(q_i^{(l)})^2) \\
&\leq \omega \sum_{l \in L}(1 - \frac{1}{K}(\sum_{i=1}^{K} q_i^{(l)})^2) \\
&= \omega \sum_{l \in L}(1 - \frac{1}{K}) \\
&= [t(M-1)+1](1 - \frac{1}{k})\omega
\end{aligned} \tag{9}$$

where ω is the heaviest partition node and the first inequality results from the fact that uniform distribution maximizes the entropy. So we get

$$\omega \geq \frac{G_t^g}{(t+1)(M-1)(1-\frac{1}{K})}$$

This gives the lower-bound on G_t^g of the following form:

$$\Delta_g^t \geq \frac{M^2\gamma^2 G_t^g}{4K(M(1-2\gamma)+2\gamma)(t+1)(M-1)(1-\frac{1}{K})} = \tilde{\Delta}_g^t$$

Let $\eta^g = \frac{2M\gamma}{\sqrt{K(M(1-2\gamma)+2\gamma)(M-1)(1-\frac{1}{k})}}$
Then we get $\tilde{\Delta}_g^t = \frac{(\eta^g)^2 G_t^g}{16(t+1)}$

Following the recursion of the proof in Section 3.2 in (Choromanska et al., 2016) (note that in our case $G_t^g \leq 2(M-1)(1-\frac{1}{K})$, we obtain that under the Weak Hypothesis Assumption, for any $k \in [0, 2(M-1)(1-\frac{1}{K})]$, to obtain $G_t^g \leq k$ it suffices to make:

$$t >= (\frac{2(M-1)(1-\frac{1}{K})}{k})^{\frac{8K[M(1-2\gamma)+2\gamma](1-\frac{1}{k})(M-1)}{M^2\gamma^2 \log_2 e}}$$

split. We can get the same resluts as Theoreme 2 by normalizing the range of k to $[0, 1]$.

We next proceed to linking Gini-entropy to errror. The multi-class classification error can be writen as (equation 20,Choromanska et al., 2017):

$$\epsilon(T) = \sum_{l \in L} \omega_l(1 - max(q_1^{(l)}, q_2^{(l)}, \cdots, q_K^{(l)}))$$

We now consider Gini-entropy, let $i_l = argmax_{i=1,2,\cdots,K} q_i^{(l)}$

$$\begin{aligned}
G_t^g &= \sum_{l \in L} \omega_l \sum_{i=1}^{K} q_i^{(l)} (1 - q_i^{(l)}) \\
&\geq \sum_{l \in L} \omega_l \sum_{i=1}^{K} q_i^{(l)} (1 - max(q_1^{(l)}, q_2^{(l)}, \cdots, q_K^{(l)})) \\
&= \sum_{l \in L} \omega_l (1 - max(q_1^{(l)}, q_2^{(l)}, \cdots, q_K^{(l)})) \sum_{i=1}^{K} q_i^{(l)} \\
&= \sum_{l \in L} \omega_l (1 - max(q_1^{(l)}, q_2^{(l)}, \cdots, q_K^{(l)})) \\
&= \epsilon(T)
\end{aligned} \tag{10}$$

Now we can say that minimizing Gini-entropy leads to reduction of multi-class classification error.

Futhermore, it is said that the behavior of the error closely mimics the behavior of the Gini-entropy (Section 4, Choromanska&Choromanski, 2016). We also provide the proof:

$$\epsilon(T) = \sum_{l \in L} \omega_l (1 - max(q_1^{(l)}, q_2^{(l)}, \cdots, q_K^{(l)})) = \sum_{l \in L} \omega_l \sum_{i=1}^{K} q_i^{(l)} (1 - max(q_1^{(l)}, q_2^{(l)}, \cdots, q_K^{(l)}))$$

$$G_t^g = \sum_{l \in L} \omega_l \sum_{i=1}^{K} q_i^{(l)} (1 - q_i^{(l)})$$

The distribution of $q_i^{(l)}$ is more likely to be uniform when there are more splits. So the difference between $1 - max(q_1^{(l)}, q_2^{(l)}, \cdots, q_K^{(l)})$ and $1 - q_i^{(l)}$ will become less and less. Thus, they have similar behavior.

Theorem 3. under the Weak Hypothesis Assumption, for any $k \in [0, 1]$, to obtain $\epsilon(T) < k$ ($\epsilon(T)$ is error of the tree T), it suffices to have a tree with $N >= (\frac{1}{k})^{\frac{8C^3 K \sqrt{KC-1}(M-1)[M(1-2\gamma)+2\gamma]}{M^2 \gamma^2 \log_2 e (C-2)^2}}$ internal nodes when we use modified Gini-entropy to estimate the quality of the tree (under the same condition as Theorem 1).

Proof

Notice that $q = \sum_{M}^{j=1} p_j q_j$. Then We denote the difference between the contribution of node n to the value of the entropy-based objectives in times t and $t + 1$ as:

$$\Delta_m^t = G_t^m - G_{t+1}^m = \omega[\tilde{G}^m(q) - \sum_{j=1}^{M} p_j \tilde{G}^m(q_j)] \tag{11}$$

Then we extend the inequality given by Equation 7 in (Choromanska et al., 2016) by applying Theorem5.2. from (Azocaretal.,2011) and obtain the following bound

$$\begin{aligned}
\Delta_m^t &= \omega[\tilde{G}^m(q) - \sum_{j=1}^{M} p_j \tilde{G}^m(q_j)] \\
&\geq \omega \frac{(C-2)^2}{C^3} \sum_{j=1}^{M} p_j ||q_j - \sum_{l=1}^{M} p_l q_l||_2^2 \\
&\geq \omega \frac{(C-2)^2}{KC^3} \sum_{j=1}^{M} p_j ||q_j - \sum_{l=1}^{M} p_l q_l||_1^2 \\
&= \omega \frac{(C-2)^2}{KC^3} \sum_{j=1}^{M} p_j (\sum_{i=1}^{K} |\frac{q_i p_{j|i}}{p_j} - \sum_{l=1}^{M} p_l \frac{q_i p_{l|i}}{p_l}|)^2 \\
&= \omega \frac{(C-2)^2}{KC^3} \sum_{j=1}^{M} \frac{1}{p_j} (\sum_{i=1}^{K} q_i |p_{j|i} - p_j|)^2
\end{aligned} \tag{12}$$

ω can further be lower-bounded by noticing the :

$$\omega \geq \frac{G_t^m}{(t+1)(M-1)\sqrt{KC-1}}$$

Then we get:

$$\Delta_m^t \geq \frac{(C-2)^2 G_t^m}{KC^3(t+1)(M-1)\sqrt{KC-1}} \sum_{j=1}^{M} \frac{1}{p_j} (\sum_{i=1}^{K} q_i |p_{j|i} - p_j|)^2$$

Notice that $p_j \leq \frac{M(1-2\gamma)+2\gamma}{M}$.

$$\begin{aligned}
\Delta_m^t &\geq \frac{(C-2)^2 G_t^m M^2}{(M(1-2\gamma)+2\gamma)KC^3(t+1)(M-1)\sqrt{KC-1}} \sum_{j=1}^{M} \frac{1}{p_j} (\sum_{i=1}^{K} q_i |p_{j|i} - p_j|)^2 \\
&\geq \frac{(C-2)^2 G_t^m M^2}{(M(1-2\gamma)+2\gamma)4KC^3(t+1)(M-1)\sqrt{KC-1}} Jn^2
\end{aligned} \tag{13}$$

where the last inequality is a consequence of Jensens inequality. Then we replace Jn as γ, and denote $\tilde{\Delta}_m^t$ as this difference:

$$\Delta_m^t \geq \frac{(C-2)^2 G_t^m M^2 \gamma^2}{(M(1-2\gamma)+2\gamma)4KC^3(t+1)(M-1)\sqrt{KC-1}} = \tilde{\Delta}_m^t$$

Let

$$\eta^m = \frac{2M\gamma}{\sqrt{\frac{C^3}{(C-2)^2} K(M(1-2\gamma)+2\gamma)(M-1)(\sqrt{KC-1})}}$$

Then we get $\tilde{\Delta}_m^t = \frac{(\eta^m)^2 G_t^m}{16(t+1)}$

Do the same trick as the proof of theorem 2, we obtain that under the Weak Hypothesis Assumption, for any $k \in [\sqrt{C-1}, 2(M-1)(\sqrt{KC-1})]$, to obtain $G_t^g \leq k$ it suffices to make:

$$t >= (\frac{2(M-1)\sqrt{KC-1}}{k})^{\frac{8C^3 K\sqrt{KC-1}(M-1)[M(1-2\gamma)+2\gamma]}{M^2\gamma^2 \log_2 e(C-2)^2}}$$

split. We can get the same resluts as Theoreme 3 by normalizing the range of k to $[0, 1]$.

We next proceed to linking modified Gini-entropy to errror.

$$G_t^m = \sum_{l \in L} \omega_l \sum_{i=1}^{K} \sqrt{q_i^{(l)}(C - q_i^{(l)})}$$

First we normalize its range, like figure 2 in Choromanska et al., 2016.

$$\dot{G}_t^m = \frac{G_t^m - \sqrt{C-1}}{\sqrt{2C-1} - \sqrt{C-1}}$$

$C > 2$, So we have $\sqrt{2C-1} - \sqrt{C-1} < \sqrt{C-1}$.

$$\begin{aligned}
\dot{G}_t^m &> \frac{\sum_{l \in L} \omega_l (\sum_{i=1}^{K} \sqrt{q_i^{(l)}(C - q_i^{(l)})} - \sqrt{C-1})}{\sqrt{C-1}} \\
&= \frac{\sum_{l \in L} \omega_l \sum_{i=1}^{K} (\sqrt{q_i^{(l)}(C - q_i^{(l)})} - q_i\sqrt{C-1})}{\sqrt{C-1}} \\
&> \frac{\sum_{l \in L} \omega_l \sum_{i=1}^{K} (\sqrt{q_i^{(l)}(C - q_i^{(l)})} - q_i\sqrt{C-q_i})}{\sqrt{C-q_i}} \\
&= \sum_{l \in L} \omega_l \sum_{i=1}^{K} (\sqrt{q_i^{(l)}} - q_i) \\
&> \sum_{l \in L} \omega_l \sum_{i=1, i \neq i_l}^{K} (\sqrt{q_i^{(l)}} - q_i)
\end{aligned} \tag{14}$$

We know that $\forall_{i=1,2,\cdots,K, i \neq i_l} q_i^l \leq 0.5$, So$\sqrt{q_i} \geq \sqrt{2} q_i$. Then we get:

$$\dot{G}_t^m > (\sqrt{2}-1) \sum_{l \in L} \omega_l \sum_{i=1, i \neq i_l}^{K} q_i = (\sqrt{2}-1)\epsilon(T)$$

Now we find the connection between modified Gini-entropy and multi-class classification error.

4. numerical experiments

We run LOMtree algorithm, which is implemented in the open source learning system Vowpal Wabbit Langford et al. (2007), on multiclass datasets, Isolet (26 classes, downloaded from http://www.cs.huji.ac.il/ shais/datasets/ClassificationDatasets.html)

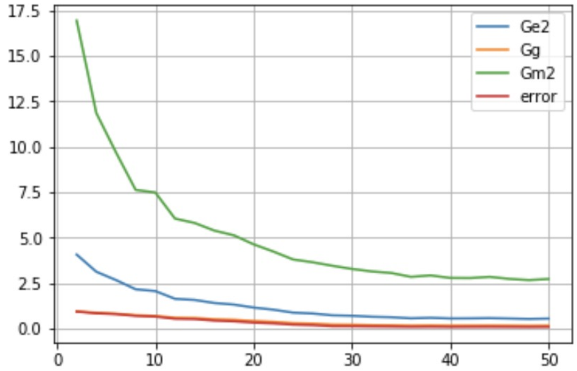

Figure 1: Isolet, Figure is recommended to be read in color.

The fisrt figure shows the relationship between Gini-entropy and modified Gini-entropy under our proof. It is clearly that error is always lower than these two entropy.

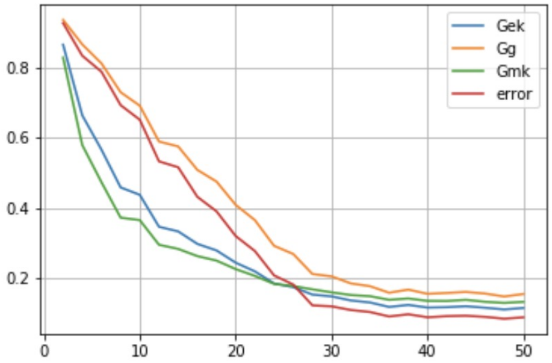

Figure 2: Isolet, Figure is recommended to be read in color.

We normalize the range of entropy-based criteria to [0,1] in the second figure. We can see that after some splits. error is going to be lower than all entropy-based criteria. So we can say that minizing entropy-based criteria leads to reduction of multi-class classification error.

5. conclusion

In this paper, we proved the error is upper-bounded by Gini entropy and its modified version after normalization. We showed that to achieve the same level of error, Gini entropy and its modified version takes more nodes when the number of classes of the input is large. We also showed that Gini entropy mimics the behavior of the error when the splits are balanced and pure.

refrences

Azocar, A., Gimenez, J., Nikodem, K., and Sanchez, J. L. On strongly mid-convex functions. Opuscula Math., 31 (1):1526, 2011

Choromanska,A.andLangford,J. Logarithmictimeonline multiclass prediction. In NIPS. 2015.

Choromanska, A., Choromanski, K., and Bojarski, M. On the boosting ability of top-down decision tree learning algorithm for multiclass classication. CoRR, abs/1605.05223, 2016.

Choromanska, A., Jernite, Y., Sontag, D. Simultaneous Learning of Trees and Representations for Extreme Classication and Density Estimation. 1610.04658, 2017

J. Langford, L. Li, and A. Strehl. http://hunch.net/ vw, 2007.

M. Kearns and Y. Mansour. On the boosting ability of top-down decision tree learning algorithms. Journal of Computer and Systems Sciences, 58(1):109128 (also In STOC, 1996), 1999