Diversity with Similarity as a Measure of Training Set Quality

based on 2407.15724

Josiah Couch

Beth Israel Deaconess Medical Center

13 August 2024

Motivation

Figure: A high diversity dataset

Figure: A low diveristy dataset

Motivation

- We want to understand training set quality

Figure: A high diversity dataset

Figure: A low diveristy dataset

Motivation

- We want to understand training set quality
  - high quality training set $\rightarrow$ high performance model

Figure: A high diversity dataset

Figure: A low diveristy dataset

Motivation

- We want to understand training set quality
  - high quality training set $\rightarrow$ high performance model
  - (We can take this as the definition))

Figure: A high diversity dataset

Figure: A low diveristy dataset

Motivation

- We want to understand training set quality
  - high quality training set $\rightarrow$ high performance model
  - (We can take this as the definition))
- In particular, what features diagnose a dataset as high quality?

Figure: A high diversity dataset

Figure: A low diveristy dataset

Motivation

- We want to understand training set quality
  - high quality training set $\rightarrow$ high performance model
  - (We can take this as the definition))
- In particular, what features diagnose a dataset as high quality?
  - Class balance and size (obviously)

Figure: A high diversity dataset

Figure: A low diveristy dataset

Motivation

- We want to understand training set quality
  - high quality training set $\rightarrow$ high performance model
  - (We can take this as the definition))
- In particular, what features diagnose a dataset as high quality?
  - Class balance and size (obviously)
  - Dataset diversity

Figure: A high diversity dataset

Figure: A low diveristy dataset

Motivation

- We want to understand training set quality
  - high quality training set $\rightarrow$ high performance model
  - (We can take this as the definition))
- In particular, what features diagnose a dataset as high quality?
  - Class balance and size (obviously)
  - Dataset diversity
    - ★ Class internal diversity

Figure: A high diversity dataset

Figure: A low diveristy dataset

Motivation

- We want to understand training set quality
  - high quality training set $\rightarrow$ high performance model
  - (We can take this as the definition))
- In particular, what features diagnose a dataset as high quality?
  - Class balance and size (obviously)
  - Dataset diversity
    - Class internal diversity
    - Overall diversity

Figure: A high diversity dataset

Figure: A low diveristy dataset

Motivation

- We want to understand training set quality
  - high quality training set → high performance model
  - (We can take this as the definition))
- In particular, what features diagnose a dataset as high quality?
  - Class balance and size (obviously)
  - Dataset diversity
    - Class internal diversity
    - Overall diversity
- How does one measure training set diversity?

Figure: A high diversity dataset

Figure: A low diveristy dataset

Motivation

- We want to understand training set quality
  - high quality training set $\rightarrow$ high performance model
  - (We can take this as the definition))
- In particular, what features diagnose a dataset as high quality?
  - Class balance and size (obviously)
  - Dataset diversity
    - Class internal diversity
    - Overall diversity
- How does one measure training set diversity?
  - The (exponential of) entropy provides a natural notion of diversity

Figure: A high diversity dataset

Figure: A low diveristy dataset

Motivation

- We want to understand training set quality
  - high quality training set $\rightarrow$ high performance model
  - (We can take this as the definition))
- In particular, what features diagnose a dataset as high quality?
  - Class balance and size (obviously)
  - Dataset diversity
    - Class internal diversity
    - Overall diversity
- How does one measure training set diversity?
  - The (exponential of) entropy provides a natural notion of diversity
  - We will use a deformation of the entropy which accounts for similarities

Figure: A high diversity dataset

Figure: A low diveristy dataset

Motivation

- We want to understand training set quality
  - high quality training set $\rightarrow$ high performance model
  - (We can take this as the definition))
- In particular, what features diagnose a dataset as high quality?
  - Class balance and size (obviously)
  - Dataset diversity
    - ★ Class internal diversity
    - ★ Overall diversity
- How does one measure training set diversity?
  - The (exponential of) entropy provides a natural notion of diversity
  - We will use a deformation of the entropy which accounts for similarities
  - To account for both class internal and overal diversity, we will consider the entropy of the joint distribution on features + classes

Figure: A high diversity dataset

Figure: A low diveristy dataset

Diversity with Similarity

lower diversity

higher diversity

Figure: Nine species of birds are less diverse than nine species drawn from across the animal family tree

Diversity with Similarity

- The exponentials of the Rényi entropies (the *Hill numbers* [1]) have long been used to quantify diversity

$$D_q(p) = e^{H_q(p)};\ H_q(p) = -\frac{1}{q-1}\log\sum_i p^q = -\frac{1}{q-1}\log\left\langle p^{q-1}\right\rangle_p \tag{1}$$

lower diversity

higher diversity

Figure: Nine species of birds are less diverse than nine species drawn from across the animal family tree

Diversity with Similarity

- The exponentials of the Rényi entropies (the *Hill numbers* [1]) have long been used to quantify diversity

$$D_q(p) = e^{H_q(p)};\ H_q(p) = -\frac{1}{q-1}\log\sum_i p^q = -\frac{1}{q-1}\log\left\langle p^{q-1}\right\rangle_p \tag{1}$$

- E.g. in ecology

lower diversity

higher diversity

Figure: Nine species of birds are less diverse than nine species drawn from across the animal family tree

Diversity with Similarity

- The exponentials of the Rényi entropies (the *Hill numbers* [1]) have long been used to quantify diversity

$$D_q(p) = e^{H_q(p)};\ H_q(p) = -\frac{1}{q-1}\log\sum_i p^q = -\frac{1}{q-1}\log\left\langle p^{q-1}\right\rangle_p \tag{1}$$

- E.g. in ecology
- However, these measures fail to capture the following intuition:

lower diversity

higher diversity

Figure: Nine species of birds are less diverse than nine species drawn from across the animal family tree

Diversity with Similarity

- The exponentials of the Rényi entropies (the *Hill numbers* [1]) have long been used to quantify diversity

$$D_q(p) = e^{H_q(p)};\ H_q(p) = -\frac{1}{q-1}\log\sum_i p^q = -\frac{1}{q-1}\log\left\langle p^{q-1}\right\rangle_p \tag{1}$$

- E.g. in ecology
- However, these measures fail to capture the following intuition:
  - Given two communities each consisting of N uniformly distributed species, if the first consists of only species from the same genus, and the second consists of species from more than one genus, the second should be more diverse!

lower diversity

higher diversity

Figure: Nine species of birds are less diverse than nine species drawn from across the animal family tree

Diversity with Similarity

- The exponentials of the Rényi entropies (the *Hill numbers* [1]) have long been used to quantify diversity

$$D_q(p) = e^{H_q(p)};\ H_q(p) = -\frac{1}{q-1}\log\sum_i p^q = -\frac{1}{q-1}\log\left\langle p^{q-1}\right\rangle_p \tag{1}$$

- E.g. in ecology
- However, these measures fail to capture the following intuition:
  - Given two communities each consisting of N uniformly distributed species, if the first consists of only species from the same genus, and the second consists of species from more than one genus, the second should be more diverse!
  - This is still ture if I replace 'genus' with some higher taxanomic unit and 'species' by some relatively lower unit (e.g. family and genus)

lower diversity

higher diversity

Figure: Nine species of birds are less diverse than nine species drawn from across the animal family tree

Diversity with Similarity (continued)

To remedy this deficit, one may define[2]:

Diversity with Similarity (continued)

To remedy this deficit, one may define[2]:

- A similarity matrix Z_{ij} whose ijth entry encodes the similarity between species i and species j

Diversity with Similarity (continued)

To remedy this deficit, one may define[2]:

- A similarity matrix Z_{ij} whose ijth entry encodes the similarity between species i and species j
- The similarity deformed Rényi entropy,

$$H_q^Z(p) = -\frac{1}{q-1} \log \left\langle (Zp)^{q-1} \right\rangle_p = -\frac{1}{q-1} \log \sum_i p_i \left(\sum_j Z_{ij} p_j \right)^{q-1} \tag{2}$$

Diversity with Similarity (continued)

To remedy this deficit, one may define[2]:

- A similarity matrix Z_{ij} whose ijth entry encodes the similarity between species i and species j
- The similarity deformed Rényi entropy,

$$H_q^Z(p) = -\frac{1}{q-1}\log\left\langle (Zp)^{q-1}\right\rangle_p = -\frac{1}{q-1}\log\sum_i p_i\left(\sum_j Z_{ij}p_j\right)^{q-1} \tag{2}$$

- The corresponding similarity sensitive diversity $D_q^Z(p) = e^{H_q^Z(p)}$

Diversity with Similarity (continued)

To remedy this deficit, one may define[2]:

- A similarity matrix Z_{ij} whose ijth entry encodes the similarity between species i and species j
- The similarity deformed Rényi entropy,

$$H_q^Z(p) = -\frac{1}{q-1}\log\left\langle (Zp)^{q-1}\right\rangle_p = -\frac{1}{q-1}\log\sum_i p_i\left(\sum_j Z_{ij}p_j\right)^{q-1} \tag{2}$$

- The corresponding similarity sensitive diversity $D_q^Z(p) = e^{H_q^Z(p)}$
- N.B. that when Z is the identity matrix, these reduce to the ordinary Rényi entropies/Hill numbers.

Diversity with Similarity (continued)

To remedy this deficit, one may define[2]:

- A similarity matrix Z_{ij} whose ijth entry encodes the similarity between species i and species j
- The similarity deformed Rényi entropy,

$$H_q^Z(p) = -\frac{1}{q-1}\log\left\langle (Zp)^{q-1}\right\rangle_p = -\frac{1}{q-1}\log\sum_i p_i\left(\sum_j Z_{ij}p_j\right)^{q-1} \tag{2}$$

- The corresponding similarity sensitive diversity $D_q^Z(p) = e^{H_q^Z(p)}$
- N.B. that when Z is the identity matrix, these reduce to the ordinary Rényi entropies/Hill numbers.
- Also, even though we defined a similarity matrix on species, w.l.o.g. we could have defined a similarity matrix on individuals.

Diversity with Similarity (continued)

To remedy this deficit, one may define[2]:

- A similarity matrix Z_{ij} whose ijth entry encodes the similarity between species i and species j
- The similarity deformed Rényi entropy,

$$H_q^Z(p) = -\frac{1}{q-1}\log\left\langle (Zp)^{q-1}\right\rangle_p = -\frac{1}{q-1}\log\sum_i p_i\left(\sum_j Z_{ij}p_j\right)^{q-1} \tag{2}$$

- The corresponding similarity sensitive diversity $D_q^Z(p) = e^{H_q^Z(p)}$
- N.B. that when Z is the identity matrix, these reduce to the ordinary Rényi entropies/Hill numbers.
- Also, even though we defined a similarity matrix on species, w.l.o.g. we could have defined a similarity matrix on individuals.

In our context (ML for medical imaging)

Diversity with Similarity (continued)

To remedy this deficit, one may define[2]:

- A similarity matrix Z_{ij} whose ijth entry encodes the similarity between species i and species j
- The similarity deformed Rényi entropy,

$$H_q^Z(p) = -\frac{1}{q-1}\log\left\langle (Zp)^{q-1}\right\rangle_p = -\frac{1}{q-1}\log\sum_i p_i\left(\sum_j Z_{ij}p_j\right)^{q-1} \tag{2}$$

- The corresponding similarity sensitive diversity $D_q^Z(p) = e^{H_q^Z(p)}$
- N.B. that when Z is the identity matrix, these reduce to the ordinary Rényi entropies/Hill numbers.
- Also, even though we defined a similarity matrix on species, w.l.o.g. we could have defined a similarity matrix on individuals.

In our context (ML for medical imaging)

- Each image is considered a unique species

Diversity with Similarity (continued)

To remedy this deficit, one may define[2]:

- A similarity matrix Z_{ij} whose ijth entry encodes the similarity between species i and species j
- The similarity deformed Rényi entropy,

$$H_q^Z(p) = -\frac{1}{q-1}\log\left\langle (Zp)^{q-1}\right\rangle_p = -\frac{1}{q-1}\log\sum_i p_i\left(\sum_j Z_{ij}p_j\right)^{q-1} \tag{2}$$

- The corresponding similarity sensitive diversity $D_q^Z(p) = e^{H_q^Z(p)}$
- N.B. that when Z is the identity matrix, these reduce to the ordinary Rényi entropies/Hill numbers.
- Also, even though we defined a similarity matrix on species, w.l.o.g. we could have defined a similarity matrix on individuals.

In our context (ML for medical imaging)

- Each image is considered a unique species
- For simplicity, the similarity is chosen to be $Z(x,y) = \exp(-d_{RMSD}(x,y))$, where d_{RMSD} is the pixel-wise root mean squared difference

Metacommunity and subcommunity

$$A_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu) \left(\sum_j Z_{ij} p(x_j, y_\mu) \right)^{q-1} \right]^{1-q} \quad (3)$$

$$B_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu) \left(\frac{\sum_j Z_{i,j} p(x_j, y_\mu)}{\sum_j Z_{ij} p(x_j) p(y_\mu)} \right)^{q-1} \right]^{1-q} \quad (4)$$

$$\Gamma_q^Z(p) = \left[\sum_i p(x_i) \left(\sum_j Z_{ij} p(x_j) \right)^{q-1} \right]^{1-q} \quad (5)$$

$$\Delta_q^Z(p) = \left[\sum_\mu p(y_\mu)^q \right]^{1-q} \quad (6)$$

Metacommunity and subcommunity

- In many situation, one may want to compare diversity of parts to that of the whole

$$A_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu)\left(\sum_j Z_{ij} p(x_j, y_\mu)\right)^{q-1}\right]^{1-q} \quad (3)$$

$$B_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu)\left(\frac{\sum_j Z_{i,j} p(x_j, y_\mu)}{\sum_j Z_{ij} p(x_j) p(y_\mu)}\right)^{q-1}\right]^{1-q} \quad (4)$$

$$\Gamma_q^Z(p) = \left[\sum_i p(x_i)\left(\sum_j Z_{ij} p(x_j)\right)^{q-1}\right]^{1-q} \quad (5)$$

$$\Delta_q^Z(p) = \left[\sum_\mu p(y_\mu)^q\right]^{1-q} \quad (6)$$

Metacommunity and subcommunity

- In many situation, one may want to compare diversity of parts to that of the whole
  - E.g. in ecology, islands vs a whole island chain

$$A_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu)\left(\sum_j Z_{ij} p(x_j, y_\mu)\right)^{q-1}\right]^{1-q} \quad (3)$$

$$B_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu)\left(\frac{\sum_j Z_{i,j} p(x_j, y_\mu)}{\sum_j Z_{ij} p(x_j) p(y_\mu)}\right)^{q-1}\right]^{1-q} \quad (4)$$

$$\Gamma_q^Z(p) = \left[\sum_i p(x_i)\left(\sum_j Z_{ij} p(x_j)\right)^{q-1}\right]^{1-q} \quad (5)$$

$$\Delta_q^Z(p) = \left[\sum_\mu p(y_\mu)^q\right]^{1-q} \quad (6)$$

Metacommunity and subcommunity

- In many situation, one may want to compare diversity of parts to that of the whole
  - E.g. in ecology, islands vs a whole island chain
  - E.g. in ML, training set vs individual classes

$$A_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu)\left(\sum_j Z_{ij} p(x_j, y_\mu)\right)^{q-1}\right]^{1-q} \quad (3)$$

$$B_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu)\left(\frac{\sum_j Z_{i,j} p(x_j, y_\mu)}{\sum_j Z_{ij} p(x_j) p(y_\mu)}\right)^{q-1}\right]^{1-q} \quad (4)$$

$$\Gamma_q^Z(p) = \left[\sum_i p(x_i)\left(\sum_j Z_{ij} p(x_j)\right)^{q-1}\right]^{1-q} \quad (5)$$

$$\Delta_q^Z(p) = \left[\sum_\mu p(y_\mu)^q\right]^{1-q} \quad (6)$$

Metacommunity and subcommunity

- In many situation, one may want to compare diversity of parts to that of the whole
  - E.g. in ecology, islands vs a whole island chain
  - E.g. in ML, training set vs individual classes
- For this purpose, one may label each individual both by species/type and by subcommunity, and consider the joint distribution [2, 3]

$$A_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu)\left(\sum_j Z_{ij} p(x_j, y_\mu)\right)^{q-1}\right]^{1-q} \quad (3)$$

$$B_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu)\left(\frac{\sum_j Z_{i,j} p(x_j, y_\mu)}{\sum_j Z_{ij} p(x_j) p(y_\mu)}\right)^{q-1}\right]^{1-q} \quad (4)$$

$$\Gamma_q^Z(p) = \left[\sum_i p(x_i)\left(\sum_j Z_{ij} p(x_j)\right)^{q-1}\right]^{1-q} \quad (5)$$

$$\Delta_q^Z(p) = \left[\sum_\mu p(y_\mu)^q\right]^{1-q} \quad (6)$$

Metacommunity and subcommunity

- In many situation, one may want to compare diversity of parts to that of the whole
  - E.g. in ecology, islands vs a whole island chain
  - E.g. in ML, training set vs individual classes
- For this purpose, one may label each individual both by species/type and by subcommunity, and consider the joint distribution [2, 3]
  - The exponential of the entropy of the joint distribution is then termed the *Alpha diversity*

$$A_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu)\left(\sum_j Z_{ij} p(x_j, y_\mu)\right)^{q-1}\right]^{1-q} \quad (3)$$

$$B_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu)\left(\frac{\sum_j Z_{i,j} p(x_j, y_\mu)}{\sum_j Z_{ij} p(x_j) p(y_\mu)}\right)^{q-1}\right]^{1-q} \quad (4)$$

$$\Gamma_q^Z(p) = \left[\sum_i p(x_i)\left(\sum_j Z_{ij} p(x_j)\right)^{q-1}\right]^{1-q} \quad (5)$$

$$\Delta_q^Z(p) = \left[\sum_\mu p(y_\mu)^q\right]^{1-q} \quad (6)$$

Metacommunity and subcommunity

- In many situation, one may want to compare diversity of parts to that of the whole
  - E.g. in ecology, islands vs a whole island chain
  - E.g. in ML, training set vs individual classes
- For this purpose, one may label each individual both by species/type and by subcommunity, and consider the joint distribution [2, 3]
  - The exponential of the entropy of the joint distribution is then termed the *Alpha diversity*
  - The exponential of the mutual information between species and subcommunity is termed (normalized) *Beta diversity*

$$A_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu)\left(\sum_j Z_{ij} p(x_j, y_\mu)\right)^{q-1}\right]^{1-q} \quad (3)$$

$$B_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu)\left(\frac{\sum_j Z_{i,j} p(x_j, y_\mu)}{\sum_j Z_{ij} p(x_j) p(y_\mu)}\right)^{q-1}\right]^{1-q} \quad (4)$$

$$\Gamma_q^Z(p) = \left[\sum_i p(x_i)\left(\sum_j Z_{ij} p(x_j)\right)^{q-1}\right]^{1-q} \quad (5)$$

$$\Delta_q^Z(p) = \left[\sum_\mu p(y_\mu)^q\right]^{1-q} \quad (6)$$

Metacommunity and subcommunity

- In many situation, one may want to compare diversity of parts to that of the whole
  - E.g. in ecology, islands vs a whole island chain
  - E.g. in ML, training set vs individual classes
- For this purpose, one may label each individual both by species/type and by subcommunity, and consider the joint distribution [2, 3]
  - The exponential of the entropy of the joint distribution is then termed the *Alpha diversity*
  - The exponential of the mutual information between species and subcommunity is termed (normalized) *Beta diversity*
  - That of the marginal distribution on species/subcommunity are termed *Gamma/Delta diversity*

$$A_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu)\left(\sum_j Z_{ij} p(x_j, y_\mu)\right)^{q-1}\right]^{1-q} \quad (3)$$

$$B_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu)\left(\frac{\sum_j Z_{i,j} p(x_j, y_\mu)}{\sum_j Z_{ij} p(x_j) p(y_\mu)}\right)^{q-1}\right]^{1-q} \quad (4)$$

$$\Gamma_q^Z(p) = \left[\sum_i p(x_i)\left(\sum_j Z_{ij} p(x_j)\right)^{q-1}\right]^{1-q} \quad (5)$$

$$\Delta_q^Z(p) = \left[\sum_\mu p(y_\mu)^q\right]^{1-q} \quad (6)$$

Metacommunity and subcommunity

- In many situation, one may want to compare diversity of parts to that of the whole
  - E.g. in ecology, islands vs a whole island chain
  - E.g. in ML, training set vs individual classes
- For this purpose, one may label each individual both by species/type and by subcommunity, and consider the joint distribution [2, 3]
  - The exponential of the entropy of the joint distribution is then termed the *Alpha diversity*
  - The exponential of the mutual information between species and subcommunity is termed (normalized) *Beta diversity*
  - That of the marginal distribution on species/subcommunity are termed *Gamma/Delta diversity*
- All of this goes through when considering a similarity on species (we won't consider a similarity on subcommunities)

$$A_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu)\left(\sum_j Z_{ij} p(x_j, y_\mu)\right)^{q-1}\right]^{1-q} \quad (3)$$

$$B_q^Z(p) = \left[\sum_{i,\mu} p(x_i, y_\mu)\left(\frac{\sum_j Z_{i,j} p(x_j, y_\mu)}{\sum_j Z_{ij} p(x_j) p(y_\mu)}\right)^{q-1}\right]^{1-q} \quad (4)$$

$$\Gamma_q^Z(p) = \left[\sum_i p(x_i)\left(\sum_j Z_{ij} p(x_j)\right)^{q-1}\right]^{1-q} \quad (5)$$

$$\Delta_q^Z(p) = \left[\sum_\mu p(y_\mu)^q\right]^{1-q} \quad (6)$$

Methodology

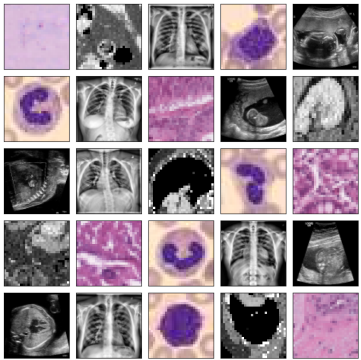

Figure: Images from some of the selected datasets

Methodology

1. Collect a number of datasets

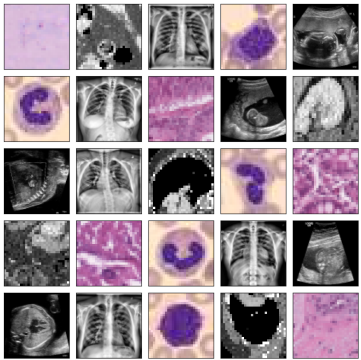

Figure: Images from some of the selected datasets

Methodology

1. Collect a number of datasets
   - PathMNIST, BloodMNIST, OrganAMNIST, and OrganCMNIST from MedMNIST [4, 5]

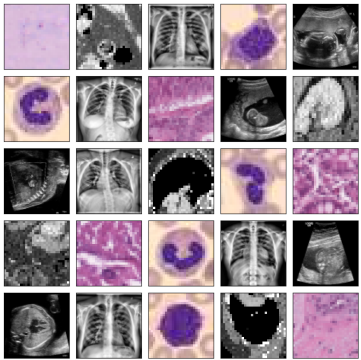

Figure: Images from some of the selected datasets

Methodology

1. Collect a number of datasets
   - PathMNIST, BloodMNIST, OrganAMNIST, and OrganCMNIST from MedMNIST [4, 5]
   - Several additional datasets, including those used in Madani et al. [6] and Chinn et al. [7]

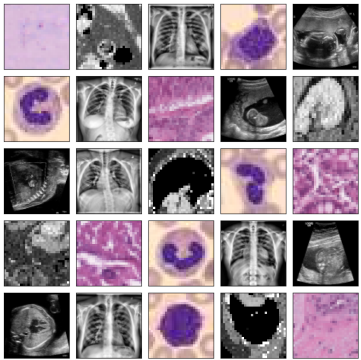

Figure: Images from some of the selected datasets

Methodology

1. Collect a number of datasets
   - PathMNIST, BloodMNIST, OrganAMNIST, and OrganCMNIST from MedMNIST [4, 5]
   - Several additional datasets, including those used in Madani et al. [6] and Chinn et al. [7]
2. From each dataset, sample the training set to create many subsets

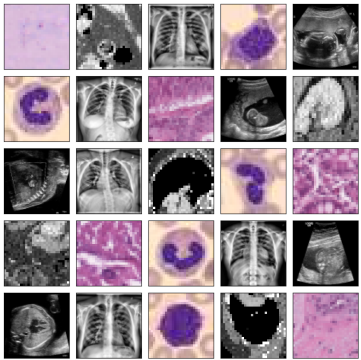

Figure: Images from some of the selected datasets

Methodology

1. Collect a number of datasets
   - PathMNIST, BloodMNIST, OrganAMNIST, and OrganCMNIST from MedMNIST [4, 5]
   - Several additional datasets, including those used in Madani et al. [6] and Chinn et al. [7]
2. From each dataset, sample the training set to create many subsets
3. Train classifier on each subset

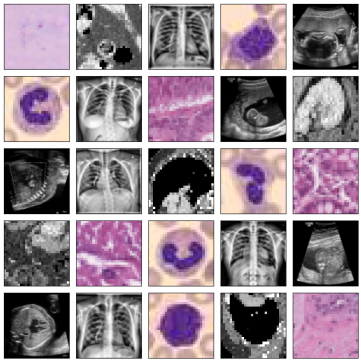

Figure: Images from some of the selected datasets

Methodology

1. Collect a number of datasets
   - PathMNIST, BloodMNIST, OrganAMNIST, and OrganCMNIST from MedMNIST [4, 5]
   - Several additional datasets, including those used in Madani et al. [6] and Chinn et al. [7]
2. From each dataset, sample the training set to create many subsets
3. Train classifier on each subset
4. Test each of these models against test set (common to subsets from a given source)

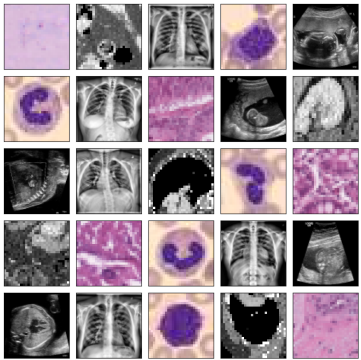

Figure: Images from some of the selected datasets

Methodology

1. Collect a number of datasets
   - PathMNIST, BloodMNIST, OrganAMNIST, and OrganCMNIST from MedMNIST [4, 5]
   - Several additional datasets, including those used in Madani et al. [6] and Chinn et al. [7]
2. From each dataset, sample the training set to create many subsets
3. Train classifier on each subset
4. Test each of these models against test set (common to subsets from a given source)
5. Measure an assortment of diversity indices for each subset (including class balance)

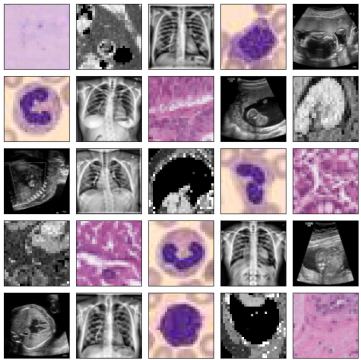

Figure: Images from some of the selected datasets

Methodology

1. Collect a number of datasets
   - PathMNIST, BloodMNIST, OrganAMNIST, and OrganCMNIST from MedMNIST [4, 5]
   - Several additional datasets, including those used in Madani et al. [6] and Chinn et al. [7]
2. From each dataset, sample the training set to create many subsets
3. Train classifier on each subset
4. Test each of these models against test set (common to subsets from a given source)
5. Measure an assortment of diversity indices for each subset (including class balance)
6. Use linear regression to measure how much variation in model performance is explained by different sets of diversity indices.

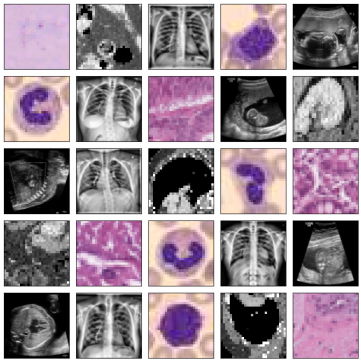

Figure: Images from some of the selected datasets

Methodology

1. Collect a number of datasets
   - PathMNIST, BloodMNIST, OrganAMNIST, and OrganCMNIST from MedMNIST [4, 5]
   - Several additional datasets, including those used in Madani et al. [6] and Chinn et al. [7]
2. From each dataset, sample the training set to create many subsets
3. Train classifier on each subset
4. Test each of these models against test set (common to subsets from a given source)
5. Measure an assortment of diversity indices for each subset (including class balance)
6. Use linear regression to measure how much variation in model performance is explained by different sets of diversity indices.
7. (Our models are actually log linear, and each model includes the image size, number of color channels, and number of classes as features)

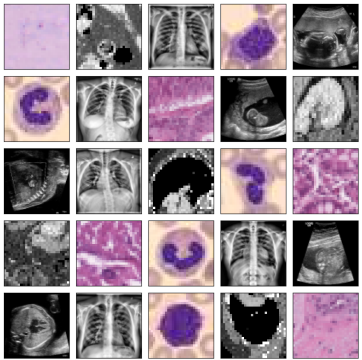

Figure: Images from some of the selected datasets

Results

We found that some diversities with similarity, especially the Alpha diversity (at Rényi parameters 0 and 1), outperform class balance, both singly and when combined with size.

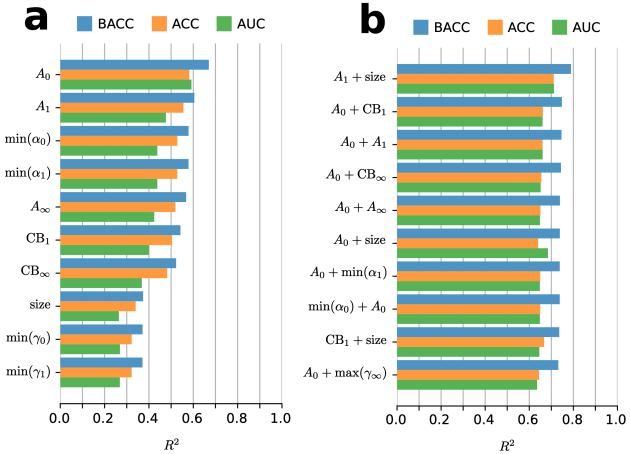

Figure: R^2 for linear models of resnet-18 accuracy trained on different sets of diversities (a) single features (b) pairs of features

Acknowlegements

I'd like to thank IAIFI for hosting this workshop, as well as my collaborators, Dr. Ramy Arnaout (BIDMC and HMS) and Dr. Rima Arnaout (UCSF Department of Medicine). This work was supported by the Gordon and Betty Moore foundation and by the NIH.

Figure: Link to paper: `https://arxiv.org/abs/2407.15724`

Thank you for attending my talk!

References

M.O. Hill, *Diversity and evenness: A unifying notation and its consequences*, *Ecology* **54** (1973) 427.

T. Leinster and C.A. Cobbold, *Measuring diversity: the importance of species similarity*, *Ecology* **93** (2012) 477.

R. Reeve, T. Leinster, C.A. Cobbold, J. Thompson, N. Brummitt, S.N. Mitchell et al., *How to partition diversity*, 2016.

J. Yang, R. Shi and B. Ni, *Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis*, in *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 191–195, 2021.

J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke et al., *Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification*, *Scientific Data* **10** (2023) 41 [2110.14795].

A. Madani, R. Arnaout, M. Mofrad and R. Arnaout, *Fast and accurate view classification of echocardiograms using deep learning*, *npj Digital Medicine* **1** (2018) 6.

E. Chinn, R. Arora, R. Arnaout and R. Arnaout, *Enrich: Exploiting image similarity to maximize efficient machine learning in medical imaging*, *Journal of the American Medical Informatics Association* **30** (2023) 1079 [medrxiv.org/content/10.1101/2021.05.22.21257645].