PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

SOBRE LA CONSTRUCCIÓN DE ENSEMBLES DE CLASIFICADORES DIVERSOS EN TANTO QUE VARIACIÓN NORMALIZADA DE INFORMACIÓN Y SU VÍNCULO CON SU PRECISIÓN

(ON DIVERSE CLASSIFIER'S ENSEMBLE BUILDING BY NORMALIZED VARIATION OF INFORMATION AND ITS LINK TO ITS ACCURACY)

TRABAJO DE INVESTIGACIÓN PARA OPTAR EL GRADO ACADÉMICO DE MAGÍSTER EN INFORMÁTICA

AUTOR

Rodrigo José Guinea Ordóñez

ASESOR

Edwin Rafael Villanueva Talavera

Mayo, 2021

RESUMEN

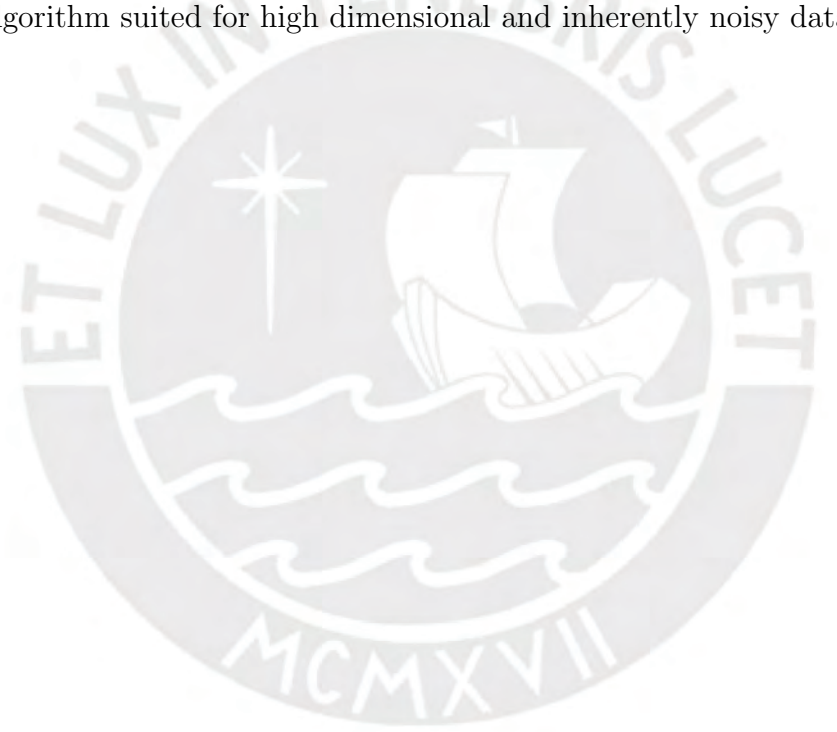
La hipótesis en cuestión afirma que, dado el contexto teórico (i.e., definiciones matemáticas consideradas apropiadas para describir los fenómenos que se pretende estudiar) descrito en el artículo, existe una relación entre diversidad global y precisión de un ensamble de clasificadores. Por lo tanto, el propósito de esta investigación es estudiar la relación entre la precisión de ensambles y su diversidad dentro de un contexto geométrico y de información. Para lograrlo, interpretamos el problema como uno geométrico introduciendo un espacio métrico, donde los puntos son predicciones de clasificadores; la función de distancia, la métrica Variación de Información Normalizada (NVI, por sus siglas en inglés); y la construcción de un ensamble diverso es reducida a un problema de criba y novedosamente transformado a uno de programación cuadrática. La significancia estadística es asegurada haciendo uso de métodos Monte Carlo sobre 53 conjuntos de datos apropiados. El resultado es un algoritmo basado en una métrica usada en el contexto de teoría de la información, ideal para estudiar conjuntos de datos de alta dimensionalidad e inherentemente ruidosos. Por tanto, es relevante cuando el costo de adquirir muestras es muy alto; y la cantidad de variables, enorme. El marco teórico incluye las definiciones (e.g., definiciones relacionadas al concepto de diversidad o al espacio métrico utilizado), los teoremas (e.g., propiedades de espacios métricos) y algoritmos base (i.e., programación cuadrática) usados para conseguir los resultados. Los resultados muestran que, en promedio, el exceso de precisión de un ensemble diverso respecto de su contraparte aleatoria es función del valor de la diversidad global del mismo. Esto confirma la hipótesis inicial. Además, la metodología introducida para modelar el algoritmo introduce un marco que permite esclarecer la relación entre diversidad y precisión, ya que la representa en términos geométricos.

Highlights

On diverse classifier's ensemble building by normalized variation of information and its link to its accuracy

Rodrigo Guinea, Edwin Villanueva

- Geometric approach to diverse ensemble building
- Relationship between accuracy and global diversity as defined as the Normalized Variation of Information
- Algorithm suited for high dimensional and inherently noisy datasets



On diverse classifier's ensemble building by normalized variation of information and its link to its accuracy

Rodrigo Guinea[a], Edwin Villanueva[b]

[a] *Graduate School, Pontifical Catholic University of Peru, Av. Universitaria 1801, San Miguel, Peru, Lima, 15088, Peru*
[b] *Graduate School, Pontifical Catholic University of Peru, Av. Universitaria 1801, San Miguel, Peru, Lima, 15088, Peru*

Abstract

Ensemble models for classification are a Machine Learning approach that have frequently proven useful in generating results with higher performance and robustness than mono-classifier models. Common advantages include tolerance for input data noise, decreased variance, and bias in predictions. Many studies justify the fact that the diversity of an ensemble is related to accuracy in some way. However, the correct definition of diversity and the conditions needed for those statements to hold true remain unclear. The present work addresses this issue from a geometrical perspective presenting a method to build diverse ensembles based on the *Normalized Variation of Information* and explore which conditions correlate to the variability in its accuracy. The knowledge generated from this analysis will make it possible to clarify and bring insight into how ensemble diversity is related to ensemble accuracy.

Keywords: ensemble learning, high dimensional datasets, bioinformatics, diversity, metric space, machine learning

1. Introduction

Ensemble methods are useful for classification tasks because of their ability to reduce noise, variance and bias. For this reason, they are frequently used when the dataset's dimensionality is elevated, such as in the Bioinformatics

Email addresses: rguinea@pucp.edu.pe (Rodrigo Guinea), ervillanueva@pucp.edu.pe (Edwin Villanueva)

context [1, 2, 3]. In such cases it is often difficult to learn the true decision frontier with a single learner. Experimental and theoretical results suggest that a single learner can be enough to classify correctly a fraction of the samples; but an ensemble of learners that err, with the correct amount of overlap in different samples, can increase the accuracy significantly [4, 5].

There are several ways in which the individual classifications can be combined [6]. One of the most popular combination strategies is the majority voting scheme, because of its simplicity, effectiveness and intuitive nature. This kind of strategy has been thoroughly studied for more than 20 years. For example, it has been shown that an ensemble of better than random independent classifiers (i.e., probability of success greater than 0.5) are bound to perform better than any individual base component [7]. Works like [8, 7, 9, 10, 11, 12] have shown that when dependence between classifiers exists, it is possible to increase the accuracy of the ensemble further still. In [7], L. I. Kuncheva calculates an upper limit for accuracy improvement in her famous *pattern of success*, given a set of classifiers and a majority voting combiner function.

An important concept associated with ensemble learning is that of diversity, which has to do with the degree of disagreement among base learners. Several measures have been proposed to quantify diversity [13]. Many researchers believe that this concept is key in ensemble construction. The thought process can be outlined as follows [14, 7, 15]:

1. There exists a property called *diversity* that is linked to the ensemble accuracy;
2. Such a property can be used to build an effective ensemble of classifiers.

Despite the acceptance of the importance of diversity, until now there is no measure of it that is widely accepted [16, 14]. There are many candidates [13, 17, 18, 19, 20] each of which gives good results in a different context or as part of different kinds of algorithms [21, 22].

Diversity measures are mainly used as: i) descriptive statistics of ensembles, ii) criteria to build ensembles procedurally [20], or iii) criteria for reducing an existing ensemble of learners, such that the resulting reduced set is diverse according to the chosen definition of diversity (task known as ensemble pruning) [23].

Ensemble pruning can be formulated as a Quadratic Programming problem. This formulation possesses attractive properties, such as its well-founded

theoretical base [24] and its wide availability of solvers (e.g. [25, 26, 27]). Previous studies defined an objective function (i.e., function to be optimized) and reformulated it in such a way to be optimized as quadratic programming problem [28, 29] (i.e., risk or margin optimization). But neither of those works use explicitly the concept of diversity for pruning. Y. Zhang et al. [30] defined a symmetric objective matrix called *ensemble error* as a function of the overall strength of ensemble classifiers and an ad hoc diversity. They arranged the objective matrix in a quadratic form to model it as a quadratic integer programming problem. In the present study we follow the quadratic programming formulation. However, differently from previous studies, we use an information based metric to directly build a positive definite dissimilarity matrix to fit in a quadratic form. By applying convex relaxation to the optimizing vector, limiting its range to [0,1] and normalizing it, the problem can be changed to a continuous optimization one. With this well-known technique, the relation between ensemble accuracy and diversity is empirically investigated in a comprehensive set of high dimensional datasets with an intensive simulation approach.

In Section 2, the mathematical framework, tools, and notation will be presented; as well as the specific problem to be studied. In Section 3, the materials will be introduced; the experimental setup, described; the results, shown; and their analysis, explained. Finally, Section 4 will conclude with some remarks and future directions this study could take.

2. Theory and Definitions

2.1. On the Majority Voting Combiner

Let $F = \{f_i\}_{i=1}^{T}$ be a set, called ensemble, of T multi-class classifiers defined over some discrete domain $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots\}$, where $\boldsymbol{x}_i$ is called a data point; and C, a vector of labels. Each element of $F(X) = \{f_i(X)\}_{i=1}^{T}$ is called a prediction and take values in $C^{|X|}$. Now, let $h_{ic}(\boldsymbol{x}) \in \{0, 1\}$ be an indicator function that returns 1 if the *ith* classifier takes the value $c \in C$ for data point $\boldsymbol{x}$; and 0, otherwise. Then, the *majority vote* combiner function on $\boldsymbol{x}$ can be defined as [31]:

$$V(\boldsymbol{x}) = \arg\max_{c \in C} \left(\sum_{i=1}^{T} h_{ic}(\boldsymbol{x}) \right) \tag{1}$$

where $V(\boldsymbol{x})$ gives us the most voted class $c \in C$, among the T classifiers, for element $\boldsymbol{x}$.

2.2. On Diversity Statistics

According to L.I. Kuncheva [13], diversity statistics can be considered to fall under two categories: pairwise and global (i.e., non-pairwise). Pairwise statistics assess diversity between two predictions. Global diversity statistics are set functions which assign a number to $F(X)$ to represent its diversity. For simplicity, f_i and f_j are said to be *diverse* on X if their respective predictions $f_i(X)$ and $f_j(X)$ are diverse. It is also assumed that if datasets X and Y comes from the same distribution, $f_i(Y)$ and $f_j(Y)$ are diverse too.

2.2.1. Normalized Variation of Information

There are many statistics used as definition for diversity in the literature [18] [13]. Nevertheless, this study adopts a pairwise metric called *Normalized Variation of Information* (*NVI*), proposed by Marina Meilă [32],

$$\begin{aligned} NVI(f_i, f_j) &= 1 - \frac{I(f_i, f_j)}{H(f_i, f_j)} \\ &\propto H(f_i \mid f_j) + H(f_j \mid f_i) \end{aligned} \tag{2}$$

where H is the entropy and I is the Mutual Information, defined for two classifiers f_i, f_j as:

$$\begin{aligned} I(f_i, f_j) &= H(f_i) - H(f_i \mid f_j) \\ &= \sum_{x \in f_i(\boldsymbol{x}), y \in f_j(\boldsymbol{x})} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

The function *NVI* was chosen because of its interpretation, which will be explained in section 2.3.2, and definition as a proper metric. That is, it satisfies the non-negativity, symmetry and triangle inequality axioms. These properties are ideal for geometrical interpretation [32], which this paper will fully exploit. However, since the dimension of X is finite, any distance between $f_i(X)$ and $f_j(X)$ will be an approximation of their true distance (i.e., the distance between predictions of two classifiers on an infinite dimensional dataset point $\boldsymbol{x}$). Yijun Bian *et al.* [33] also used *NVI* for ensemble pruning and interpreted it as a measure of redundancy between predictions, since when two classifiers are maximally diverse (i.e., distance equal to 1), they

are minimally redundant. And the other way around; if they are maximally redundant, they will be minimally diverse (i.e., distance 0). In this work, however, *NVI* will be treated directly as a diversity metric. From (2), it follows that the diversity between two predictions $f_i(X)$ and $f_j(X)$ is proportional to the sum of the uncertainty in $f_i(X)$ that is left after considering the information gained by $f_j(X)$, and the uncertainty that is left in $f_j(X)$ that is left after considering the information gained by $f_i(X)$. In other words, *NVI* measures the sum of uncertainties of $f_i(X)$ and $f_j(X)$ after subtracting the information that flows from one to the other.

2.2.2. NVI Global Diversity

Skalak [34] defined *global diversity* as the average of all pairwise *disagreement measures* [35] of an ensemble base classifiers. But in this study, we define global diversity using the pairwise metric instead,

$$G = \frac{2}{T(T-1)} \sum_{i=1}^{T} \sum_{j=i+1}^{T} NIV(f_i, f_j) \tag{3}$$

where *NVI* is given by (2); and T, the number of classifiers in the ensemble.

2.3. On Diverse Ensemble Building

The problem of diverse ensemble building will be solved by pruning and modeled into quadratic programming. So, given a set F of T classifiers defined on some domain X, consider their predictions $F(X) \in \mathcal{F}$ as points in a metric space $(\mathcal{F}, div, X)$ with distance div. Now, let $div(f_i, f_j) = d_{ij}$ be the pair-wise distances between any two points $f_i(X)$ and $f_j(X)$ and $\boldsymbol{D} = (d_{ij})$ a dissimilarity matrix. To build a diverse ensemble a subset from $\mathcal{F}$ of maximally diverse points is selected. This problem can be initially formulated as that of a combinatorial optimization problem. Let $w = (w_1, \ldots, w_T)$ be a binary valued vector, were its *ith* element takes the value of 1 if f_i is selected; and 0, otherwise. Then, the problem would be to maximize

$$M(\boldsymbol{w}) = \boldsymbol{w}^t \boldsymbol{D} \boldsymbol{w} \tag{4}$$

subject to $\sum_i w_i = n$, where n is the number of points to be selected. This is the classical optimization problem known to be *NP* hard. However, convex relaxation can be applied to reduce the combinatorial problem into a continuous optimization one [36]. This can be done by allowing the vector w_i

take values in the interval $[0, 1]$. Additionally, $\sum_i w_i = 1$ to prevent unlimited growth of the norm during the optimization process. In this way, the problem is solved by

$$\boldsymbol{w_{max}} = \arg\max_{\boldsymbol{w}} M(\boldsymbol{w}) \tag{5}$$

Since D is a positive definite symmetric matrix, $M(\boldsymbol{w})$ is convex, and thus a global optimum can be found [37]. There are many quadratic programming solvers and some of them have polynomial runtimes [25, 26, 27, 38]. With this formulation, the initial problem of selecting the k maximally distant points is not solved directly. However, now this result can be interpreted geometrically. Each *ith* component in $\boldsymbol{w}$ can be thought as a weight w_i that measures how diverse (i.e., far from every other point) is $f_i(X)$ in $\mathcal{F}$. This can be seen by expanding (4) into

$$M(\boldsymbol{w}) = \sum_{i<j\leq T} w_i w_j d_{i,j} \tag{6}$$

Each point $f_i(X)$ is associated with one and only one w_i. Since all terms in (6) are non negative, a point $f_i(X)$ will be assigned a larger weight than some $f_k(X)$ if the sum of distances between $f_i(X)$ and the rest of the points is greater than that of $f_k(X)$. In other words, if w_i is found consistently in terms with larger distances than w_k, then w_i will be larger that w_k. In this study, the weights will be arbitrarily rounded to two decimal places in order to drop any sufficiently redundant classifier (i.e., when it's prediction weight is approximately zero).

2.3.1. A Proof of concept with probability of disagreement metric

As a first proof of concept, the *Probability of Disagreement* will be used as the definition of diversity. Given a dataset $X = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_N})$, the probability of disagreement is defined for two predictions $f_i(X)$ and $f_j(X)$ as [39]:

$$\begin{aligned} d_{i,j} &= P(f_i(X) \neq f_j(X)) \\ &= \frac{1}{N}\sum_{k=1}^{N}\left(1 - \delta_{f_i(x_k)f_j(x_k)}\right) \\ &= 1 - \frac{1}{N}\sum_{k=1}^{N}\delta_{f_i(x_k)f_j(x_k)} \end{aligned} \tag{7}$$

where δ is the Kronecker delta. This diversity statistic ensures a metric space where the triangular inequality, symmetry and non-negativity holds. However, since $f_k(X) < \infty$, (7) is only an approximation. Just as (2), (7) can be used as a distance in a geometrical space. Also, from its definition, it is easy to prove that if the classifiers map into a set of $|C|$ classes, then there could only simultaneously exist $|C|$ maximally distant predictions. That is, $|C|$ predictions that are at distance 1 from each other. This fact will be used to model an appropriate example.

Consider six classifiers $\{f_i\}_{i=1}^{6}$ and a set of labels $C = \{1, 2, 3\}$. Their respective predictions are shown in (8). Note that classifier's f_4, f_5, f_6 predictions are maximally diverse. Also, assume an hypothetical ground truth $\boldsymbol{c} = (2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3)$ (i.e., the true labeling of X).

$$\begin{aligned}
f_1(X) &= (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 3) \\
f_2(X) &= (1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 3, 3, 1) \\
f_3(X) &= (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 3, 1, 1) \\
f_4(X) &= (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1) \\
f_5(X) &= (2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2) \\
f_6(X) &= (3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3)
\end{aligned} \tag{8}$$

By the use of (7) the distance matrix, defined as $\boldsymbol{D} = (d_{ij})$, is

$$\boldsymbol{D} = \begin{bmatrix}
0.00 & 0.31 & 0.15 & 0.15 & 1.00 & 0.85 \\
0.31 & 0.00 & 0.31 & 0.31 & 0.85 & 0.85 \\
0.15 & 0.31 & 0.00 & 0.15 & 0.92 & 0.92 \\
0.15 & 0.31 & 0.15 & 0.00 & 1.00 & 1.00 \\
1.00 & 0.85 & 0.92 & 1.00 & 0.00 & 1.00 \\
0.85 & 0.85 & 0.92 & 1.00 & 1.00 & 0.00
\end{bmatrix} \tag{9}$$

After maximizing the corresponding quadratic form (5), the resulting vector of weights is

$$\boldsymbol{w_{max}} = (0.00, 0.01, 0.00, 0.33, 0.33, 0.33) \tag{10}$$

In order to gain some intuition from these results, a classic multidimensional scaling method (i.e., principal coordinates analysis) [40] was used to find an embedding that represents the classifiers predictions as points in a

two-dimensional space. This method tries to preserve the distance between points as much as possible.

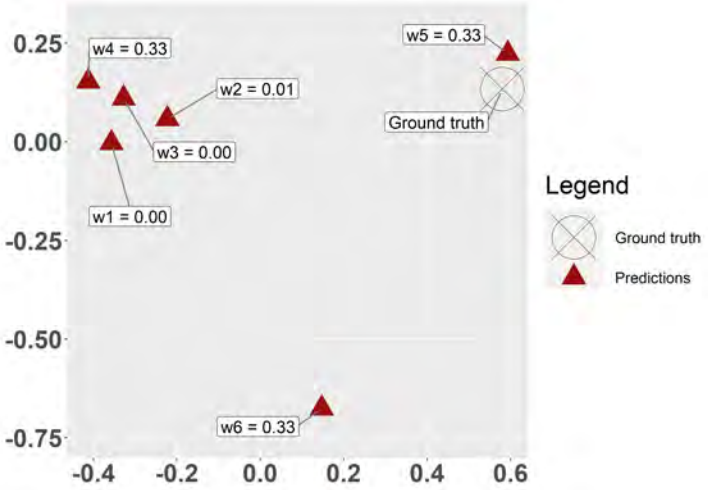


Figure 1: Two dimensional representation of the distance matrix (9) by principal coordinates analysis. This plot includes an hypothetical ground truth.

The resulting embedding (Figure 1), shows that the algorithm assigns a greater weight to the predictions that are more distant from each other, as expected. f_4, f_5 and f_6 had been assigned a weight of 0.33. f_1 and f_3 are really close to f_4 (i.e., redundant with respect to f_4), so they have been assigned a weight 0. f_2, although close to f_4, is far enough to have been assigned a non-zero weight of 0.01.

2.3.2. A Proof of concept with Normalized Variation of Information metric

As in the previous case, six classifiers were considered. This time, the classifiers were obtained from training a random forest ensemble of 6 decision trees on the Iris dataset [41]. The resulting distance matrix is shown in (11). Unfortunately, in this case, no pairs of classifiers have maximum distance.

$$\boldsymbol{D} = \begin{bmatrix} 0.00 & 0.27 & 0.10 & 0.18 & 0.18 & 0.24 \\ 0.27 & 0.00 & 0.21 & 0.15 & 0.20 & 0.27 \\ 0.10 & 0.21 & 0.00 & 0.10 & 0.13 & 0.24 \\ 0.18 & 0.15 & 0.10 & 0.00 & 0.18 & 0.18 \\ 0.18 & 0.20 & 0.13 & 0.18 & 0.00 & 0.25 \\ 0.24 & 0.27 & 0.24 & 0.18 & 0.25 & 0.00 \end{bmatrix} \tag{11}$$

Solving the quadratic programming problem (5) gives:

$$\boldsymbol{w_{max}} = (0.25, 0.29, 0.00, 0.00, 0.16, 0.30) \tag{12}$$

Although, the classifier predictions are not shown explicitly as in 2.3.1, it is important to note that a vector representation of those is not appropriate when using information based measures, such as NVI. Because of the way those are defined [32], each prediction must be seen as a vector of sets of distinguishable elements. That is, a prediction is now modeled as a vector of a particular partition $f(X) = (\{a_1, a_2, a_3, a_4\}, \{a_5, a_6\})$, instead of $f(X) = (a_1, a_2, a_3, a_4, a_5, a_6)$. This is precisely the reason why NVI was chosen as the definition of diversity in this study. For example, consider a pair of hypothetical predictions $f_1(X) = (1, 1, 1, 2, 2)$ and $f_2(X) = (2, 2, 2, 1, 1)$. The probability of disagreement distance would be $d_{1,2} = 1$. On the other hand, $NVI(f_1, f_2) = 0$. Moreover, consider

$$\begin{aligned} f_1 &= (\{a_1, a_2, a_3\}, \{a_4, a_5\}) \\ f_2 &= (\{a_4, a_5\}, \{a_1, a_2, a_3\}) \\ f_3 &= (\{a_1, a_2, a_5\}, \{a_3, a_4\}) \end{aligned} \tag{13}$$

Then, $NVI(f_1, f_2) = 0$, but $NVI(f_1, f_3) = 0.9896$. The diversity captured by NVI is about the underlying patterns in the dataset that leads a model to establish a particular partition. For a more concrete exemplification, think about persons as classifiers. Consider a situation where a person f_k must judge (i.e., classify) a set of phenomena $X = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_5})$ into two different categories $\mathcal{C}_1$ and $\mathcal{C}_2$. That is, $f_k(X) = (\mathcal{C}_1, \mathcal{C}_2)$, where $\mathcal{C}_1$ represents the subset of elements in $f_k(X)$ judged as *right*; and $\mathcal{C}_2$, as wrong. In particular, f_1 could have judged $\mathcal{C}_1 = \{a_1, a_2, a_3\}, \mathcal{C}_2 = \{a_4, a_5\}$; and f_2, $\mathcal{C}_1 = \{a_4, a_5\}, \mathcal{C}_2 = \{a_1, a_2, a_3\}$. Although both have an equivalent partition, they are *morally* opposites. Since the partitions were the same, it can be said their knowledge (i.e., information)

allowed them to form the same categories, although named differently. Any discussion between them would solely be about the naming of those categories.

This changes the way of thinking about *closeness* in the metric space defined by NVI. Now two classifiers being far away from each other means that they group the elements of the dataset into very different partitions. That is, the partition of one does not reduce the uncertainty (i.e., cannot be used as a reference) about the partition of the other. Even if the labels are 'wrongly' assigned, the information given by $f_2(X)$ completely determines $f_1(X)$, and vice versa.

In Figure 2 that $w_3 = w_4 = 0$. This is because $f_3(X)$ is surrounded by $f_1(X)$ and $f_5(X)$; and $f_4(X)$, by $f_2(X)$, $f_5(X)$ and $f_6(X)$. In other words, by knowing the information provided by $f_3(X)$, for example, more is known about the partition of $f_5(X)$ and $f_1(X)$. Thus, $f_3(X)$ is redundant and can be discarded for the sake of diversity.

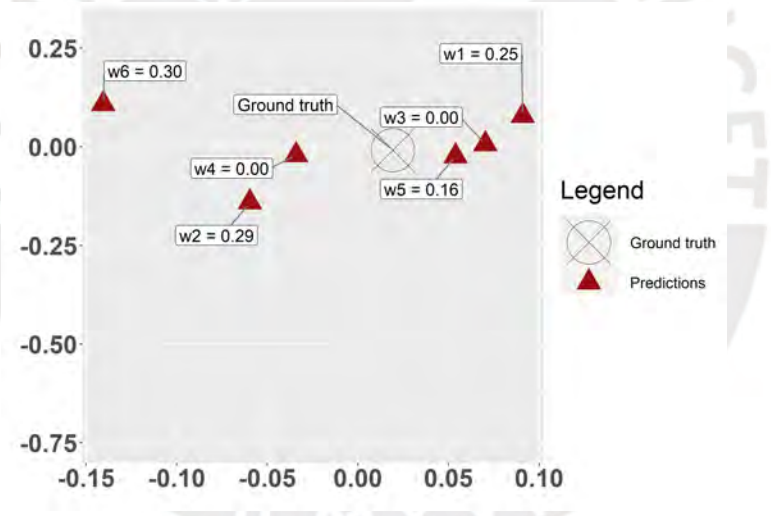


Figure 2: Two dimensional representation of the distance matrix (11) by principal coordinates analysis. This is the representation of a space where the NVI represents the distance between two classifiers.

2.4. On Controlling Bias

To account for the ground truth in the pruning process, a class relevance term is added to (4). Classifiers near the ground truth will be given more weight (i.e., less penalized during optimization). The new quadratic programming formulation is reflected in Equation 14 where the second term accounts

for the ground truth:

$$M(\boldsymbol{w}, k) = \boldsymbol{w}^\top \boldsymbol{D}\boldsymbol{w} - C_d(k)^\top \boldsymbol{w} \tag{14}$$

The vector $C_d(k)$ has T components, one for each classifier, in which its ith component represents the distance between prediction $f_i(X)$ and the ground truth $\boldsymbol{c}$ elevated to the power of k. In this way, the optimization procedure will avoid giving too much weight to those classifiers that are far from $\boldsymbol{c}$. Also, classifiers that are farther from c are penalized more than those that are closer. It is important to consider that function (14) can be reduced to (4) when $k \to \infty$.

As before, we first use the probability of disagreement as a proof of concept. The maximization will be done on (14) with $k = 1$. Assuming $\boldsymbol{c} = (2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3)$ is the ground truth and by solving (5), the resulting vector of weights is

$$\boldsymbol{w_{max}} = (0.00, 0.00, 0.00, 0.18, 0.63, 0.19) \tag{15}$$

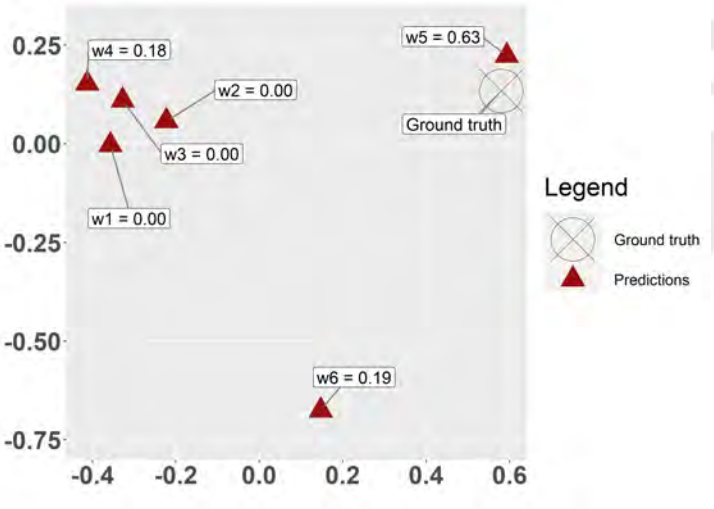


Figure 3: Two dimensional representation of the distance matrix (9) by principal coordinates analysis. This plot includes an hypothetical ground truth.

Figure 3 shows the resulting 2D embedding. The weight assigned to f_2 is now 0, since it has been penalized because its distance from the ground truth $\boldsymbol{c}$ is large enough. The remaining weight has been redistributed in such a way that f_5 now has 0.63, f_4 has 0.18 and f_6 has 0.19 weight.

3. Experimentation and Results

3.1. Materials

In this study 49 microarray and 4 RNASeq high dimensional datasets were used. 39 out of the 49 microarray datasets are described in A. Espichan and E. Villanueva 2018 study [2]. They have two classes, positive or negative cancer prognostic status, and at most 2000 variables each. The average, minimal and maximal number of samples is 87, 40 and 286, respectively. As for the rest of the microarray datasets, they were fetched from the Gene Expression Ómnibus (GEO).

Regarding the 4 RNAseq, because of computational power limitations, we only select the first 2000 variables from each. All of them (see Table 1) are imbalanced and contain two classes, pathological tissue (PT) and normal tissue (NT).

Table 1: RNASeq datasets

Dataset	**Class**	**Samples**
BLCA	PT	408
	NT	19
HNSC	PT	520
	NT	44
KIRC	PT	533
	NT	72
PRAD	PT	497
	NT	52

3.2. Experimental Setup

In this section, a metric space $(\mathcal{F}, NVI, X)$ is defined. $\mathcal{F}$ represents a set of points; NVI, the distance; and X, the dataset (domain) used by an algorithm to generate the points in $\mathcal{F}$. As for the algorithm, Random Forest [42] is used to build an ensemble of 100 decision trees on a training set (i.e., subset of X) comprised by 75% of the data points in X. Solving (5) together with (14), and rounding w to three decimal places, a subset of classifiers is selected. This subset will be called a *diverse ensemble*. Additionally, a

random subset, the same size as the diverse ensemble, is selected from $\mathcal{F}$ and is called *baseline ensemble*. To compare the performance of the diverse ensemble against the baseline ensemble, the accuracy of both is computed on a test set, comprised by the remaining 25% data points in X, with respect to the ground truth. That is, the accuracy of the former is divided against the latter, and the result is subtracted by 1. The result is the *increment*[1] of accuracy of the diverse ensemble with respect to the baseline. An increment of 0 would mean that both ensembles result in the same accuracy.

To account for uncertainty, a Monte Carlo approach is taken. Thus, this procedure is repeated 1500 times for each one of the 53 datasets considered in this study, resulting in as many distributions of increments (i.e., increment of the accuracy of the diverse ensemble with respect to the baseline). During this process, other distributions were computed for each dataset. Those were: the percentage of the classifiers that were pruned, the global diversity of the baseline ensemble and the global diversity of the diverse ensemble.

Since some distributions included few, but extreme, outliers, the trimmed mean at 5% is used as the central value statistic in this study. Nevertheless, plots shown in section 3.3 can be found in the supplementary material for trimmed means at 0%, 10%, 15%, 20% and 25%.

3.3. Results and discussion

Using the results from the simulations outlined in 3.2, it can be seen in Figure 4 that the hypothesised trade-off between the mean of the mean accuracy of the ensemble's base classifiers and the global diversity is confirmed [30].

[1] In this study, the *increment* of any statistic will always be with respect to the baseline ensemble.

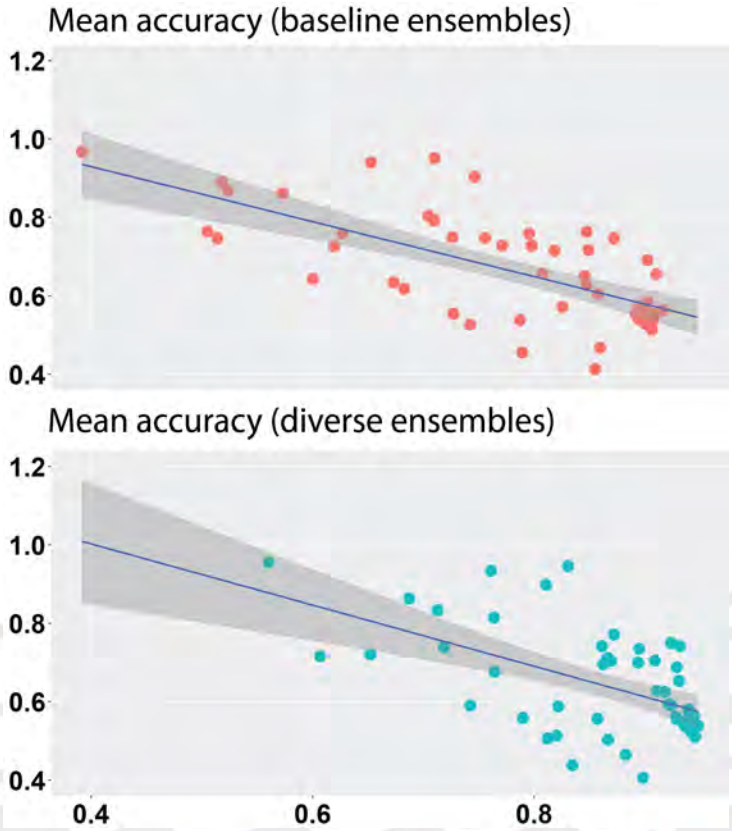


Figure 4: For the case in which $k \to \infty$, the vertical axis in each plot represents the mean accuracy of the base classifiers within each ensemble. The horizontal axis represents the mean global diversity of their respective ensembles. The p-values for both, the intercept and coefficient, for the baseline and diverse ensembles were at most $1.86e-05$.

The relation between the distributions of mean diversities for each hyper-parameter k were compared using histograms (see Figure 5).

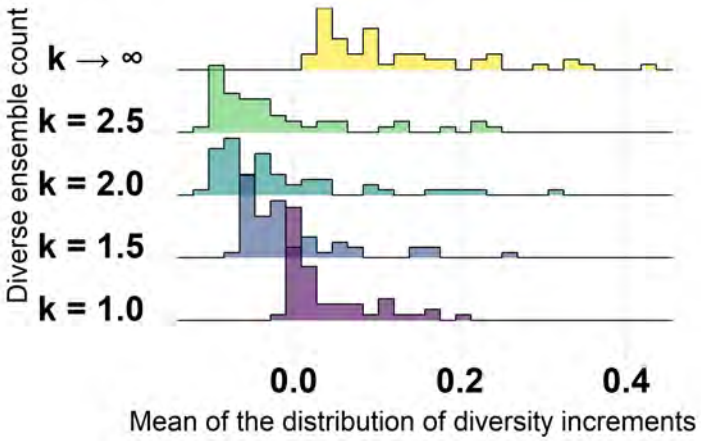


Figure 5: The vertical axis represents the count of diverse ensembles build from their respective datasets, each one associated with a distribution of diversity increments. The horizontal axis represents the mean of the distribution of diversity increments.

As expected, every dataset from the case in which $k \to \infty$ is more diverse than the baseline, as it was also expected to exist a trade off between diversity and class relevance. Nevertheless, it was also the case that 68.63% and 26.42% of the ensembles for $k = 1.0$ and $k = 1.5$, respectively, were more diverse than the baseline, while for $k = 2.0$, 32.08. Meaning that it is not necessarily the case that the more influence from the ground truth is considered, using the class relevance term in 14, the less diverse the resulting ensembles will be.

Similarly, in Figure 6, the performance, across datasets, of the diverse ensembles tend to perform better when $k = 2$ rather than when $k < 2$, even if in the latter case the ensembles were built with a stronger class relevance influence. The mean and median for $k = 1.0$ is 0.00510 and 0.00178; $k = 1.5$ is 0.02140 and 0.00673; for $k = 2.0$, 0.02589 and 0.00868; and for $k \to \infty$, −0.01461 and −0.00953, respectively.

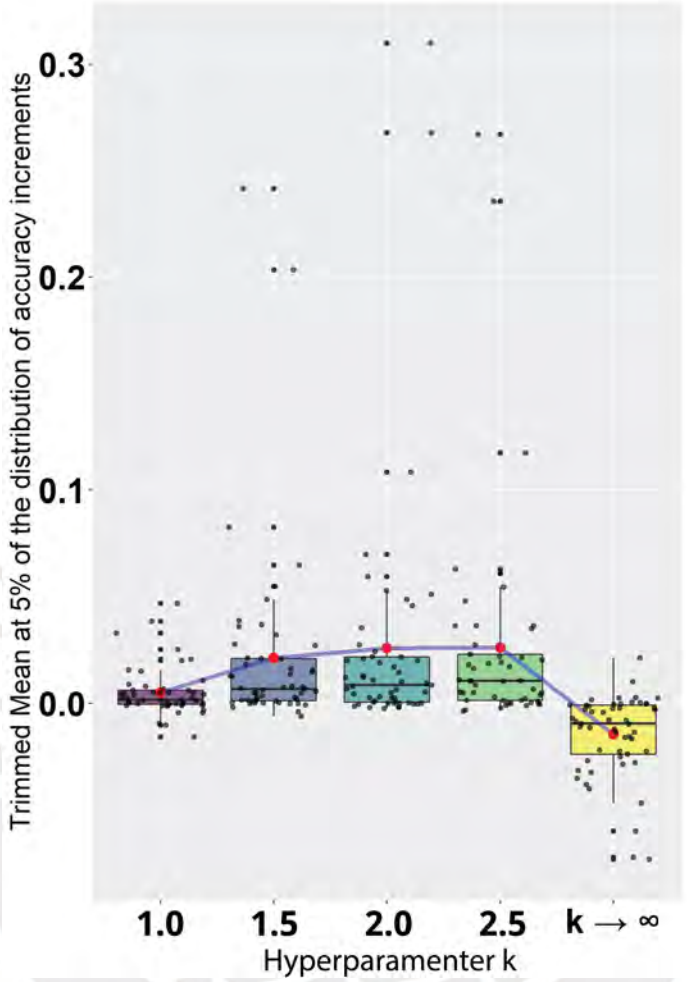


Figure 6: Each black point represent a dataset. The vertical axis represent the mean of the distribution of accuracy increments with respect to its baseline ensemble. The horizontal axis represent the values for the k hyperparameter. The greater the value of k, the less impact will have the class relevance term. Each red dot represent the mean accuracy of all datasets for each k.

Figure 6 shows that diversity alone does not guarantee good accuracy, since the worst results were obtained when $k \to \infty$ (no influence of ground truth). It also suggests that a bigger influence of class relevance does not necessarily improve results either.

Figure **??** depicts in more detail the relationship between diversity and the distribution of accuracy increments. This plot is consistent with 6 in that

diversity alone does not give good results. It also suggests that for $k = 2.0$, the diverse ensembles perform better than for $k = 1.5$. Additionally, the more diverse the baseline ensemble, the more difference does it make to build a diverse ensemble using (14). The standard deviation increases with diversity, which can be worrying at first, but then ones notice that the skewness becomes increasingly positive too, giving a heavy tail on the positive side.

To understand this results, each concept used in the experimental setup must be inspected carefully. Random forest fits weak learners on subsets of the dataset (i.e., datasets formed by randomly selecting subset of variables and samples). Because the baseline ensemble is built by randomly choosing classifiers from the random forest ensemble, their diversity is approximately the same (see Figure 8). A low diversity of the resulting baseline ensemble, implies that most base learners are redundant. That is, it does not matter which subset is learned, the base learner will give similar predictions. Hence, we can say the dataset is redundant. This can also be interpreted as the decision boundary is *easy* to learn. On the other side, if the baseline ensemble shows high diversity, the dataset is non redundant, as well as its base learners.

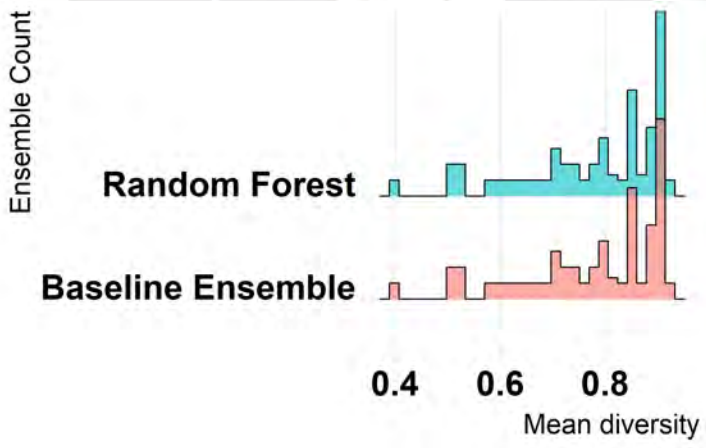


Figure 8: This figure compare the histograms of mean diversities, across datasets, of random forest ensembles against baseline ensembles. The overlap is clear, even if the number of counts of the baseline is less than the number of counts from the random forest.

Now, since the random forest is a supervised learning algorithm and if the dataset is easy to learn, we can expect ensembles with high accuracy (see Figure **??**) and with predictions relatively close to each other, and closely clustered symmetrically around the ground truth. Figure 9 shows a two

dimensional projection of one particular simulation of the distribution of classifications. This distribution results from fitting a random forest to the *KIRC* RNASeq dataset and building a diverse and baseline ensembles ($k \to \infty$). The average of the distances between every base classifier from the random forest, baseline and diverse ensemble to the ground truth is 0.5212, 0.5090 and 0.6071. These results are consistent with Figure 9, in view of the fact that the predicted values from the random forest and baseline ensemble are placed in the same position. The skewness of those same distances for the random forest, baseline and diverse ensemble is −0.0683, 0.1578, −0.3541, respectively. This shows that the classifications are fairly symmetrical with respect to the ground truth.

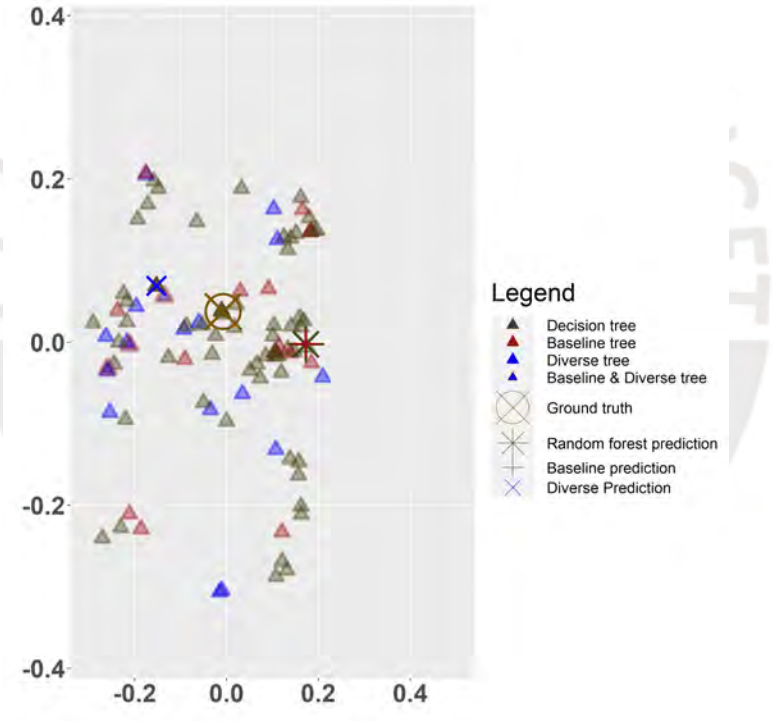


Figure 9: This figure shows an embedded two dimensional spacial distribution of the classifiers fitted on the low diversity KIRC RNASeq dataset. The embedding was done with principal coordinates analysis. The hyperparameter, $k \to \infty$ was used to build the diverse ensemble..The random forest, baseline and diverse ensemble's diversity is 0.5868, 0.5669354 and 0.7072188. The the accuracy of all three ensembles is 0.9868.

If the diversity is high (i.e., the dataset is hard to learn), we expect a

lower accuracy (see Figure **??**), and the base classifiers to not be clustered around the ground truth, but skewed (because of systematic error). Figure 10 corroborates this hypothesis, which corresponds to the results on the *pancreas Ishikawa* microarray dataset. In this case, the average of the distances between every base classifier from the random forest, baseline and diverse ensemble to the ground truth is 0.9577, 0.9537 and 0.9653. All of them are similar and much higher than the previous example, suggesting that they are farther from the ground truth. Also, after the random forest ensemble was pruned, the resulting diverse and baseline ensemble distributions of predictions around the ground truth remained similar. But, at the same time, the accuracy from the random forest, baseline and diverse ensemble was 0.8462, 0.8462 and 0.6154. The diverse ensemble performs much lower than the other ones. A possible explanation would be that, since subsets of the dataset tend to be non redundant (i.e., contain different information), classifier's predictions do not tend to share information among themselves. As a result, removing some of them disregarding the amount of information they contain about the true classification ($k \to \infty$) will probably reduce the accuracy. Finally, and contrary to the previous example, the skewness of the distribution of distances to the ground truth of each ensemble is much larger (i.e., −2.0074, −1.7344 and −2.1701 for the random forest, baseline and diverse ensemble, respectively). This skewness can be seen in Figure 10.

For statistical significance, the previous examples will be repeated 500 times to show that the tendencies illustrated by Figures 9 and 10 are the norm rather than the exception. For the high diversity example, the mean and the standard deviation of the skewness in the random forest, baseline, and diverse ensemble are −2.9457 and 1.0507, −2.6039 and 0.9719, and −2.3009 and 0.8167, respectively. For the low diversity example, the mean and the standard deviation of the random forest, baseline and diverse ensemble are −0.3575 and 0.2197, −0.2831 and 0.3920, and −0.3861 and 0.4617, respectively. Hence, we can be confident that the examples shown in Figures 9 and 10 are representative.

4. Conclusion

This paper introduced an easy to use and intuitive new ensemble pruning method based on an entropy-based metric to study how diversity and class relevance might affect accuracy in decision trees ensembles. The results show, under the experimental context outlined in section (3.2), that diversity,

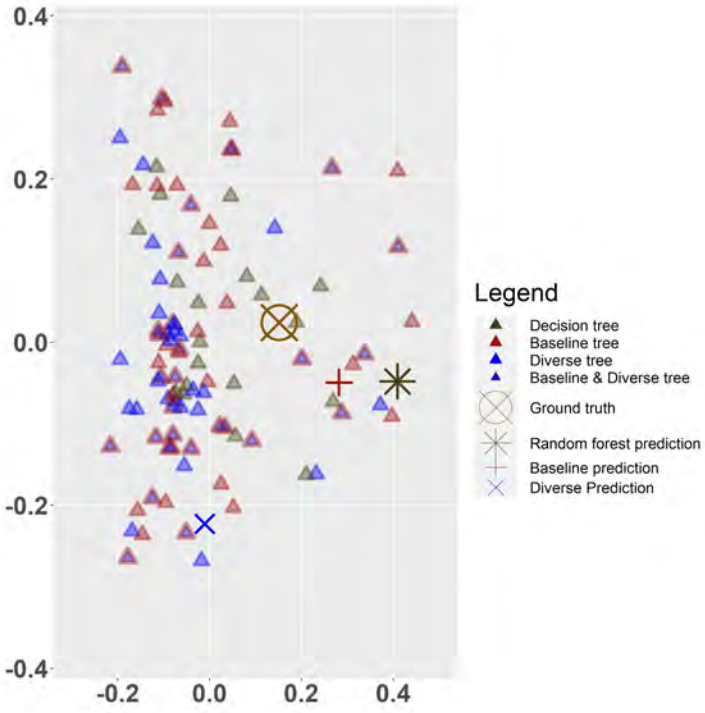


Figure 10: This figure shows an embedded two dimensional spacial distribution of the classifiers fitted on the high diversity Pancreas Ishikawa microarray dataset. The embedding was done with principal coordinates analysis. The hyperparameter, $k \to \infty$ was used to build the diverse ensemble. The random forest, baseline and diverse ensemble's diversity is 0.9569, 0.9550 and 0.9565. The accuracy of the random forest, baseline and diverse ensemble is 0.8462, 0.8462 and 0.6154, respectively.

as defined in section 2.2.1, alone is not enough to improve the ensemble performance when compared to its baseline (i.e., random forest ensemble of the same size). But it also shows that the accuracy of a diverse ensemble tends not to be in direct relation to the influence of the class relevance term (hyperparameter k), but it suggests there might be optima. A diverse ensemble, without the influence of the class relevance term (i.e., $k \to \infty$), tends to perform worst than its baseline when the diversity of $\mathcal{F}$, which is nearly equal to the diversity of the baseline, is high (i.e., near to 1). On the other side, if the diversity of $\mathcal{F}$ is low (i.e., the dataset is easy to learn), the baseline and diverse ensembles have a very similar performance. This study also shed some light on how the distribution of predictions around the ground truth is affected by diversity and how high diversity tends to

generate a systematic error, which tends to be accentuated in the $k \to \infty$ diverse ensemble case, in the predictions of the classifiers generated by $\mathcal{F}$. Futures work could use different types of base classifiers (e.g. SVM, NN or Naive Bayes). The use of strong learners would complement this paper's findings , since their properties are not the same.

References

[1] H. Yu, J. Ni, An improved ensemble learning method for classifying high-dimensional and imbalanced biomedicine data, IEEE/ACM Transactions on Computational Biology and Bioinformatics 11 (4) (2014) 657–666. doi:10.1109/tcbb.2014.2306838.
URL https://doi.org/10.1109/tcbb.2014.2306838

[2] A. Espichan, E. Villanueva, A novel ensemble method for high-dimensional genomic data classification, in: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2018. doi:10.1109/bibm.2018.8621386.
URL https://doi.org/10.1109/bibm.2018.8621386

[3] H. Mamitsuka, Empirical evaluation of ensemble feature subset selection methods for learning from a high-dimensional database in drug design, in: Third IEEE Symposium on Bioinformatics and Bioengineering, 2003. Proceedings., IEEE Comput. Soc. doi:10.1109/bibe.2003.1188959.
URL https://doi.org/10.1109/bibe.2003.1188959

[4] X. Dong, Z. Yu, W. Cao, Y. Shi, Q. Ma, A survey on ensemble learning, Frontiers of Computer Science 14 (2) (2019) 241–258. doi:10.1007/s11704-019-8208-z.
URL https://doi.org/10.1007/s11704-019-8208-z

[5] L. Breiman, Bagging predictors, Machine Learning 24 (2) (1996) 123–140. doi:10.1007/bf00058655.
URL https://doi.org/10.1007/bf00058655

[6] D. Opitz, R. Maclin, Popular ensemble methods: An empirical study, J. Artif. Int. Res. 11 (1) (1999) 169–198.

[7] L. Kuncheva, C. Whitaker, C. Shipp, R. Duin, Limits on the majority vote accuracy in classifier fusion, Pattern Analysis & Applications 6 (1)

(2003) 22–31. doi:10.1007/s10044-002-0173-7.
URL https://doi.org/10.1007/s10044-002-0173-7

[8] Z. Duan, L. Wang, K-dependence bayesian classifier ensemble, Entropy 19 (12) (2017) 651. doi:10.3390/e19120651.
URL https://doi.org/10.3390/e19120651

[9] G. Brown, J. Wyatt, Negative correlation learning and the ambiguity family of ensemble methods (2003) 266–275 doi:10.1007/3-540-44938-8_27.
URL https://doi.org/10.1007/3-540-44938-8_27

[10] L. Lam, Classifier combinations: Implementations and theoretical issues, in: Multiple Classifier Systems, Springer Berlin Heidelberg, 2000, pp. 77–86. doi:10.1007/3-540-45014-9_7.
URL https://doi.org/10.1007/3-540-45014-9_7

[11] B. E. ROSEN, Ensemble learning using decorrelated neural networks, Connection Science 8 (3-4) (1996) 373–384. doi:10.1080/095400996116820.
URL https://doi.org/10.1080/095400996116820

[12] Y. Liu, X. Yao, Ensemble learning via negative correlation, Neural Networks 12 (10) (1999) 1399–1404. doi:10.1016/s0893-6080(99)00073-8.
URL https://doi.org/10.1016/s0893-6080(99)00073-8

[13] L. I. Kuncheva, C. J. Whitaker, Machine Learning 51 (2) (2003) 181–207. doi:10.1023/a:1022859003006, [link].
URL https://doi.org/10.1023/a:1022859003006

[14] L. Didaci, G. Fumera, F. Roli, Diversity in classifier ensembles: Fertile concept or dead end?, in: Multiple Classifier Systems, Springer Berlin Heidelberg, 2013, pp. 37–48. doi:10.1007/978-3-642-38067-9_4.
URL https://doi.org/10.1007/978-3-642-38067-9_4

[15] Zhou, Ensemble methods : foundations and algorithms, Taylor & Francis, Boca Raton, FL, 2012.

[16] W. Li, R. Paffenroth, Optimal ensembles for deep learning classification: Theory and practice, in: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), IEEE, 2019.

doi:10.1109/icmla.2019.00271.
URL https://doi.org/10.1109/icmla.2019.00271

[17] M. Lakshmanamurti, Coefficient of association between two attributes in statistics, Proceedings of the Indian Academy of Sciences - Section A 22 (3) (Sep. 1945). doi:10.1007/bf03170918.
URL https://doi.org/10.1007/bf03170918

[18] L. Kuncheva, Ten measures of diversity in classifier ensembles: limits for two classifiers, in: DERA/IEE Workshop Intelligent Sensor Processing, IEE, 2001. doi:10.1049/ic:20010105.
URL https://doi.org/10.1049/ic:20010105

[19] R. E. Banfield, L. O. Hall, K. W. Bowyer, W. Kegelmeyer, Ensemble diversity measures and their application to thinning, Information Fusion 6 (1) (2005) 49–62. doi:10.1016/j.inffus.2004.04.005.
URL https://doi.org/10.1016/j.inffus.2004.04.005

[20] P. Zyblewski, M. Woźniak, Clustering-based ensemble pruning and multi-stage organization using diversity, in: Lecture Notes in Computer Science, Springer International Publishing, 2019, pp. 287–298. doi:10.1007/978-3-030-29859-3_25.
URL https://doi.org/10.1007/978-3-030-29859-3_25

[21] H. Hasanpour, R. G. Meibodi, K. Navi, Optimal selection of ensemble classifiers using particle swarm optimization and diversity measures, Intelligent Decision Technologies 13 (1) (2019) 131–137. doi:10.3233/IDT-190354.
URL https://doi.org/10.3233/IDT-190354

[22] P. Zyblewski, M. Woźniak, Novel clustering-based pruning algorithms, Pattern Analysis and Applications 23 (3) (2020) 1049–1058. doi:10.1007/s10044-020-00867-8.
URL https://doi.org/10.1007/s10044-020-00867-8

[23] G. D. Cavalcanti, L. S. Oliveira, T. J. Moura, G. V. Carvalho, Combining diversity measures for ensemble pruning, Pattern Recognition Letters 74 (2016) 38–45. doi:10.1016/j.patrec.2016.01.029.
URL https://doi.org/10.1016/j.patrec.2016.01.029

[24] N. V. Thoai, Solution methods for general quadratic programming problem with continuous and binary variables: Overview, in: Advanced Computational Methods for Knowledge Engineering, Springer International Publishing, 2013, pp. 3–17. doi:10.1007/978-3-319-00293-4_1.
URL https://doi.org/10.1007/978-3-319-00293-4_1

[25] B. Turlach, A. Weingessel, Quadprog: Functions to solve quadratic programming problems (01 2013).

[26] H. J. Ferreau, C. Kirches, A. Potschka, H. G. Bock, M. Diehl, qpOASES: a parametric active-set algorithm for quadratic programming, Mathematical Programming Computation 6 (4) (2014) 327–363. doi:10.1007/s12532-014-0071-1.
URL https://doi.org/10.1007/s12532-014-0071-1

[27] W. Huyer, A. Neumaier, MINQ8: general definite and bound constrained indefinite quadratic programming, Computational Optimization and Applications 69 (2) (2017) 351–381. doi:10.1007/s10589-017-9949-y.
URL https://doi.org/10.1007/s10589-017-9949-y

[28] N. Li, Z.-H. Zhou, Selective ensemble under regularization framework, in: Multiple Classifier Systems, Springer Berlin Heidelberg, 2009, pp. 293–303. doi:10.1007/978-3-642-02326-2_30.
URL https://doi.org/10.1007/978-3-642-02326-2_30

[29] C. Shen, H. Li, Boosting through optimization of margin distributions, IEEE Transactions on Neural Networks 21 (4) (2010) 659–666. doi:10.1109/tnn.2010.2040484.
URL https://doi.org/10.1109/tnn.2010.2040484

[30] Y. Zhang, S. Burer, W. N. Street, Ensemble pruning via semi-definite programming (special topic on machine learning and optimization), Journal of Machine Learning Research 7.
URL http://www.jmlr.org/papers/volume7/zhang06a/zhang06a.pdf

[31] S. Mao, J.-W. Chen, L. Jiao, S. Gou, R. Wang, Maximizing diversity by transformed ensemble learning, Applied Soft Computing 82 (2019) 105580. doi:10.1016/j.asoc.2019.105580.
URL https://doi.org/10.1016/j.asoc.2019.105580

[32] M. Meilă, Comparing clusterings—an information based distance, Journal of Multivariate Analysis 98 (5) (2007) 873–895. doi:10.1016/j.jmva.2006.11.013.
URL https://doi.org/10.1016/j.jmva.2006.11.013

[33] Y. Bian, Y. Wang, Y. Yao, H. Chen, Ensemble pruning based on objection maximization with a general distributed framework, IEEE Transactions on Neural Networks and Learning Systems 31 (9) (2020) 3766–3774. doi:10.1109/tnnls.2019.2945116.
URL https://doi.org/10.1109/tnnls.2019.2945116

[34] D. B. Skalak, The sources of increased accuracy for two proposed boosting algorithms, in: In Proc. American Association for Arti Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop, 1996, pp. 120–125.

[35] E. K. Tang, P. N. Suganthan, X. Yao, An analysis of diversity measures, Machine Learning 65 (1) (2006) 247–271. doi:10.1007/s10994-006-9449-2.
URL https://doi.org/10.1007/s10994-006-9449-2

[36] S. IZUMINO, N. NAKAMURA, Maximization of quadratic forms expressed by distance matrices, Hokkaido Mathematical Journal 35 (3) (2006) 641–658. doi:10.14492/hokmj/1285766422.
URL https://doi.org/10.14492/hokmj/1285766422

[37] S. A. Vavasis, Complexity theory: Quadratic programming, in: Encyclopedia of Optimization, Springer US, pp. 304–307. doi:10.1007/0-306-48332-7_65.
URL https://doi.org/10.1007/0-306-48332-7_65

[38] H. W. Borchers, pracma: Practical numerical math functions version 2.3.3 from cran (Jan 2021).
URL https://rdrr.io/cran/pracma/

[39] B. Schölkopf, J. Platt, T. Hofmann, Learnability and the doubling dimension, 2007, pp. 889–896.

[40] J. C. GOWER, Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika 53 (3-4) (1966) 325–338. doi:10.1093/biomet/53.3-4.325.
URL https://doi.org/10.1093/biomet/53.3-4.325

[41] D. F. Andrews, A. M. Herzberg, Iris data, in: Springer Series in Statistics, Springer New York, 1985, pp. 5–8. doi:10.1007/978-1-4612-5098-2_2.
URL https://doi.org/10.1007/978-1-4612-5098-2_2

[42] A. Liaw, M. Wiener, Classification and regression by randomforest, R News 2 (3) (2002) 18–22.
URL https://CRAN.R-project.org/doc/Rnews/