A Diversity-aware Model for Majority Vote Ensemble Accuracy

Nick Jin Sean Lim **Robert John Durrant**
Department of Mathematics and Statistics, University of Waikato, Hamilton, New Zealand

Abstract

Ensemble classifiers are a successful and popular approach for classification, and are frequently found to have better generalization performance than single models in practice. Although it is widely recognized that 'diversity' between ensemble members is important in achieving these performance gains, for classification ensembles it is not widely understood which diversity measures are most predictive of ensemble performance, nor how large an ensemble should be for a particular application. In this paper, we explore the predictive power of several common diversity measures and show – with extensive experiments – that contrary to earlier work that finds no clear link between these diversity measures (in isolation) and ensemble accuracy instead by using the ρ diversity measure of Sneath and Sokal as an estimator for the dispersion parameter of a Polya-Eggenberger distribution we can predict, independently of the choice of base classifier family, the accuracy of a majority vote classifier ensemble ridiculously well. We discuss our model and some implications of our findings – such as diversity-aware (non-greedy) pruning of a majority-voting ensemble.

1 Background

Ensembles of diverse classifiers are a successful and popular approach for classification, and are frequently found to have better generalization performance than single models in practice. Empirical results have shown that the accuracy of an ensemble classifier tends to increase with increasing ensemble size, with the accuracy of the individual classifiers, and with increasing diversity (where 'diversity' is variously defined and measured) between the ensemble members. However, although the effect of diversity on regression ensembles is well-understood, through the bias-variance-covariance decomposition of Brown et al. (2005), the actual relationship between these aspects of model selection and the accuracy of a *classifier* ensemble is mostly unknown.

There are many modelling choices to be made in constructing a classifier ensemble and, in particular, it is well-known in practice that the choice of both the diversity generator and of the combination method can greatly affect the overall accuracy of the ensemble learner. For classifier ensembles, of the many possible combination methods the most studied, as well as one of the most commonly used, is majority voting. In a majority vote ensemble, each base classifier in the ensemble estimates a class label, and the class label chosen by the greatest number of classifiers is selected as the output of the ensemble. In this paper we consider only majority voting as a combination scheme.

Now, in earlier literature on majority vote ensemble classifiers there have been attempts to model the ensemble accuracy based on a binomial model, that is by assuming that each classifier 'votes' for a class independently of any other and with the same probability of being correct. If these assumptions were true then by invoking the Condorcet Jury Theorem (CJT) (Condorcet, 1785) one could show that – provided each ensemble member is correct more often than not – as the ensemble size grows the ensemble becomes increasingly accurate on average, eventually being correct with probability 1 in the limit. However, this clearly does not take into account the diversity of the classifiers in the ensemble (Lam and Suen, 1997; Whitaker and Kuncheva, 2003; Kuncheva et al., 2003), in spite of the empirical evidence showing that diversity is important to the ensemble accuracy. Furthermore, the conditions on CJT are impossible to verify in practice and, taken together, these facts make it challenging to optimize any accuracy-diversity trade-off for the ensemble using a binomial model. To put this another way, such theory is weak in the sense

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

that it ignores aspects of the problem that are known empirically to be very important.
On the other hand, CJT and its generalizations have retained their fascination in various fields of the Social Sciences and Economics. In particular, works by Ladha (1995) and Berg (1993) have both proposed that the accuracy of a majority-voting system can be modelled using a Polya-Eggenberger distribution (a generalization of the well-known Beta-Binomial model, which allows for a limited range of negative-valued shape parameters). While this model is still fairly restrictive – in particular, we will still assume identical probabilities that the individual classifiers vote correctly[1] – it will allow us to model dependencies between the votes of the classifiers in the ensemble and, as we will shortly see, this will make our model accurate enough to be useful.

1.1 Roadmap

In section 2 we describe the Polya-Eggenberger (P-E) model and explain how the different ingredients in a classifier ensemble correspond to its parameters. Next, we show that a particular diversity measure, namely the ρ diversity measure of Sneath and Sokal (1963), can be interpreted directly as an estimator for the dispersion parameter of the P-E model. We continue by evaluating our model experimentally, in terms of its ability to predict the generalization performance of an ensemble classifier on simulated and real-world data, against the accuracy predictions using several different common measures of ensemble diversity. Finally, we discuss the benefits and limitations of this approach, and some potential future directions.

2 Polya-Eggenberger Distribution

The Polya-Eggenberger Distribution describes the expected number of successes in N trials drawing from the Polya urn model. In the basic Polya urn model, we have a black balls (successes) and b white balls (failures) in an urn. One ball is drawn randomly from the urn and the colour of the ball is observed. The ball is returned to the urn, and a further ball of the same colour is added. Thus, in the following draw it is more likely that the same outcome is repeated. A generalization of this model is to add $s \geq 1$ balls of the same colour as the last observation with a similar effect. The Polya-Eggenberger model generalizes the Polya urn model to allow non-integer a and b, and negative-valued s (Feller, 2008; Sen and Mishra, 1996). The definition of the P-E distribution depends on these 4 parameters, a, b, s, N, and for completeness we give it below:

Definition 1 (Polya-Eggenberger Distribution (Sen and Mishra, 1996))**.** *Let N be the number of trials in a Polya urn model, let the initial number of black balls be a and the initial number of white balls be b. Let the number of additional balls of the same colour to be added following an observation (of black or white) be s. Define $p := \frac{a}{a+b}$ and $\psi := \frac{s}{a+b}$.*

Let S_N be the number of black balls drawn after N trials. Then S_N follows a Polya-Eggenberger distribution with:

Case 1: $\psi \geq -\frac{1}{N}$

$$Pr\{S_N = k\} = \frac{\binom{-\frac{p}{\psi}}{k}\binom{-\frac{1-p}{\psi}}{n-k}}{\binom{-\frac{1}{\psi}}{n}}$$

Case 2: $\psi = 0$

$$Pr\{S_N = k\} = \binom{n}{k} p^k (1-p)^{n-k}$$

With $\binom{x}{y}$ defined for any real x and integer y as

$$\binom{x}{y} := \frac{(x)(x-1)\dots(x-y+1)}{y!}$$

Note that $\binom{x}{y}$ can also be written as $\frac{\Gamma(x+1)}{\Gamma(x-y+1)\Gamma(y+1)}$ when $x \geq y$. $\binom{x}{y} = (-1)^y\binom{-x+y-1}{y} = (-1^y)\frac{\Gamma(y-x)}{\Gamma(-x)\Gamma(y+1)}$ when $x < y$ where $\Gamma(x)$ is the Gamma function.

The parameter ψ in the P-E model can be interpreted as quantifying the increased likelihood that the next observation will be the same colour as the current observation. In other words, if p_i is the probability that the i-th observation is a success, then $p_{i+1} = \frac{p_i+\psi}{1+\psi}$ if p_i was a success and $p_{i+1} = \frac{p_i}{1+\psi}$ otherwise.
We note that for $\psi > 0$ this distribution can also be viewed as a beta-binomial distribution with $\alpha = \frac{p}{\psi}$ and $\beta = \frac{1-p}{\psi}$, when $\psi = -\frac{1}{N}$ exactly it is equivalent to a hypergeometric distribution (sampling without replacement), and when $\psi = 0$ it is equivalent to a binomial distribution. In the context of classifier ensembles we can identify the ensemble size with the number of trials N, a as the number of classifiers in the ensemble correctly labelling an example, b as the number incorrectly labelling the same example, and ψ

[1] At this point it is customary to invoke the memory of the late George Box, FRS, and remind ourselves that 'All models are wrong, but some are useful'. In fact, we will also give a heuristic argument for why this particular assumption is likely not *too* wrong in practice.

as a measure of classifier diversity - given the vote of the i-th classifier, how much more (or less) likely is it the j-th classifier will vote the same way? In the context of a majority vote ensemble, under our model the expected number of ensemble members giving the correct vote can be shown to be:
Case 1: For odd N, $\sum_{i=(N+1)/2}^{N} \sum P(S_N = i)$
Case 2: For even N, $\sum_{i=(N/2)+1}^{N} \sum P(S_N = i) + \frac{1}{2}P(S_N = N/2)$

For brevity, we will refer to the distribution defined in Definition 1 as $\text{PE}(N, p, \psi)$ or $\text{PE}(N, a, b, \psi)$ depending on which parameterization is more convenient.

3 Diversity Measures

There are many measures used to assess the diversity of a classifier ensemble. Here, we focus on the average diversity measure ρ of Sneath and Sokal (1963). We first show how this diversity measure corresponds to the parameter ψ in the definition of the Polya-Eggenberger distribution when the classifiers each have the same accuracy. Although this latter assumption is false in general, we argue that it is not *too* wide of the mark in practice: In particular, if any ensemble member has (training, or hold-out) accuracy far lower than the other ensemble members it will typically be pruned, while if any single classifier is far more accurate than the others we would likely use it by itself rather than train an ensemble. Moreover, PAC theory says that if all ensemble members have the same VC-dimension, for an i.i.d sample of fixed size n then the gap between their empirical accuracies is typically only $\mathcal{O}(1/\sqrt{n})$. Therefore, it is plausible, and in our experience true in practice, that different ensemble members tend to have similar accuracy on a given sample.

Now define $\hat{P}_{ij}$ to be the observed proportion of training observations both classifier i and j classified correctly, and $\hat{P}_i$ and $\hat{P}_j$ as the observed proportion of training observations that classifier i and classifier j classified correctly, respectively.

We can then rewrite the 2×2 contingency table for a pair of classifiers D_i and D_j (Table 1) in terms of $\hat{P}_i$, $\hat{P}_j$ and $\hat{P}_{ij}$.

Table 1: 2×2 contingency table for the classifiers D_i and D_j

	D_j Correct	D_j Wrong
D_i Correct	$\hat{P}_{ij}$	$\hat{P}_i - \hat{P}_{ij}$
D_i Wrong	$\hat{P}_j - \hat{P}_{ij}$	$1 - \hat{P}_i - \hat{P}_j + \hat{P}_{ij}$

Suppose that $\forall k \in [1, N]$, $\hat{P}_k \neq 0$. The sample diversity measure of Sneath and Sokal (1963) is then:

$$\begin{aligned}\hat{\rho}_{ij} &= \frac{\hat{P}_{ij}(1 - \hat{P}_i - \hat{P}_j + \hat{P}_{ij}) - (\hat{P}_i - \hat{P}_{ij})(\hat{P}_j - \hat{P}_{ij})}{\sqrt{\hat{P}_i\hat{P}_j(1 - \hat{P}_i)(1 - \hat{P}_j)}} \\ &= \frac{\hat{P}_{ij} - \hat{P}_i\hat{P}_j}{\sqrt{\hat{P}_i\hat{P}_j(1 - \hat{P}_i)(1 - \hat{P}_j)}}\end{aligned}$$

Therefore, $\hat{\rho}_{ij}$ is a sample estimate of the correlation between the *errors* of classifiers i and j.
Now, let $\hat{r}$ be the average of the $\hat{\rho}_{ij}$ over all pairs $i \neq j$ in the ensemble:

$$\begin{aligned}\hat{r} &:= \frac{1}{N}\frac{1}{N-1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{\hat{P}_{ij} - \hat{P}_i\hat{P}_j}{\sqrt{\hat{P}_i\hat{P}_j(1 - \hat{P}_i)(1 - \hat{P}_j)}} \\ &= \frac{1}{N}\frac{1}{N-1}\left(\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{\hat{P}_{ij}}{\sqrt{\hat{P}_i\hat{P}_j(1 - \hat{P}_i)(1 - \hat{P}_j)}} - \right. \\ &\qquad \left. \sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{\hat{P}_i\hat{P}_j}{\sqrt{\hat{P}_i\hat{P}_j(1 - \hat{P}_i)(1 - \hat{P}_j)}}\right) \qquad (1)\end{aligned}$$

If we now assume that $\forall i, j : \hat{P}_i = \hat{P}_j = \bar{p}$, the sample average classifier accuracy, equation 1 simplifies to:

$$\hat{r} = \frac{1}{N^2 - N}\frac{1}{\bar{p}(1-\bar{p})}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\left(\hat{P}_{ij} - \bar{p}^2\right) = \frac{(\overline{P_{11}}/\bar{p}) - \bar{p}}{1 - \bar{p}}$$

where $\overline{P_{11}} = \frac{1}{N^2-N}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\hat{P}_{ij}$. Writing $\overline{P_{11}}$ in terms of $\hat{r}$ and $\bar{p}$, we have:

$$\begin{aligned}\overline{P_{11}} &= \bar{p}(\hat{r}(1-\bar{p}) + \bar{p}) \\ &= \bar{p}(\hat{r} - \bar{p}\hat{r} + \bar{p}) \\ &= \bar{p}((1-\hat{r})\bar{p} + \hat{r}) \\ &= \bar{p}\frac{\bar{p} + \frac{\hat{r}}{1-\hat{r}}}{\frac{1}{1-\hat{r}}} \\ &= \bar{p}\frac{\bar{p} + \frac{\hat{r}}{1-\hat{r}}}{1 + \frac{\hat{r}}{1-\hat{r}}}\end{aligned}$$

Letting $\psi = \frac{\hat{r}}{1-\hat{r}}$, $p = \bar{p}$ we can solve for the parameters of the Polya-Eggenberger model $\text{PE}(\text{N}, \bar{p}, \frac{\hat{r}}{1-\hat{r}})$.

3.1 Other estimators of diversity

As noted before, other diversity measures could be used to estimate the parameter s or ψ in principle. An extensive survey by Kuncheva and Whitaker (2003) summarizes ten different diversity measures, several of which we also consider here. They consider the predictive power (with respect to ensemble accuracy) of different diversity measures by themselves, unlike our approach of using the diversity measure *as a parameter* in our P-E model. Contrary to our findings they conclude that (for a majority-voting ensemble) "any answers to (the) questions: 'Is there a measure that is

best for the purposes of developing committees that minimize error', and 'How can we use the measures in designing the classifier ensemble?', can only be speculative".

In much of the ensemble classification literature (e.g. Malmasi and Dras (2015); Whalen and Pandey (2013); Yang (2011)), the Yule's Q-statistic (Yule, 1900) is used to measure the diversity of classifiers in an ensemble. An appealing property of the Q-statistic is it has an intuitive interpretation as an odds-ratio and is specifically designed for discrete counts (Kuncheva et al., 2000). We found however, that this diversity measure tends to overestimate the agreement between ensemble members and can be widely off the mark when used with our Polya-Eggenberger model. Indeed, the study of Kuncheva and Whitaker (2003) concludes that – with respect to the Q-statistic –"There is no realistic framework for benchmarking classifier ensembles with either synthetic or real data", again unlike our findings using the Sneath and Sokal (1963) diversity measure as a parameter in our P-E model.

One shortcoming of both the Sneath and Sokal (1963) diversity measure and Yule (1900) Q-statistics is that we need to know beforehand (or estimate) the individual classifier accuracies before we can determine the diversity. It may be useful to be able to estimate the diversity of the classifiers independently of ensemble member accuracies, such as when evaluating diversity generation schemes. In both Ladha (1995) and Berg (1993), the authors used the cosine similarity $\frac{\boldsymbol{y}_i^T \boldsymbol{y}_j}{\|\boldsymbol{y}_i\| \|\boldsymbol{y}_j\|}$ to estimate the diversity measure ρ, where $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ are n-dimensional vectors representing the output labels on the sample of classifiers i and j respectively. While we agree that voting agreement is a natural and intuitive way to derive the increased likelihood that two voters would vote similarly, we found that this measure also tends to be even more conservative than the Q-statistic and overestimates the agreement between ensemble members. We suspect this is because the accuracies of individual ensemble members may depend on the class labels.

Inspired by the feature stability measures used for feature selection in Nogueira and Brown (2015), we also evaluate the Jaccard similarity index as an estimator for ρ. We found in practice that the Jaccard similarity index gave good estimates for our synthetic cases but poor estimates on the real-world test cases.

Overall since Sneath and Sokal's diversity measure has a natural fit to our P-E model, and also gave the best predictive performance, we will focus on this measure in the sequel. In the next section 4 we report results for Sneath and Sokal's ρ and these other measures, as well as for a binomial model, to give an indication of the size of the gap in model performance.

4 Empirical Evaluation of Model

In this section we evaluate our model empirically on synthetic and real-world datasets. Our main focus is on modelling ensemble accuracy and so we consider measures of classifier diversity as tools for estimating this. In all experiments we consider predictor vectors $x \in \mathbb{R}^d$ and labels $y \in \{-1, 1\}$ and the classifier is provided with a training set of n predictor-label pairs $T_n = \{(x_i, y_i)\}_{i=1}^n$.

Note that our model does not depend directly on the family of functions ('Hypothesis space') from which a classifier is chosen, nor on the dimensionality of the problem. To keep things manageable we consider only two diversity generators, namely random subspace (RS) (i.e. sampling $k < d$ features without replacement) – which is arguably the most common diversity-generating approach used in practice – and random projection (RP) (i.e. projecting the predictors onto a random subspace of dimension $k < d$), which is less commonly used but has recently been shown to work well in the setting where $n < d$ (Durrant, 2013). In both cases the diversity between classifiers is generated by the random choice of the k predictors available to the ensemble member for training.[2].

4.1 Experiments using Synthetic Datasets

We set $d = 1000$ to be the dimensionality of our data, $\boldsymbol{u} = (1, 0, \ldots, 0)^T$ and draw n observations $\boldsymbol{t}_i \sim (0, N(0, I_{d-1})^T / \|\boldsymbol{t}_i\|$. Note that $\boldsymbol{u}$ is orthogonal to $\boldsymbol{t}_i$. We then generate a random orthogonal matrix $\boldsymbol{R}$ by sampling its entries $R_{i,j} \sim \frac{1}{\sqrt{d}} N(0, I_d)$ and using Gram-Schmidt orthogonalization on the columns. For each $\theta = \{80°, 85°, 87.5°\}$, we then let $\boldsymbol{h} = \boldsymbol{u}\boldsymbol{R}$, and $\boldsymbol{x}_i = \cos\theta \boldsymbol{u} + \sin\theta \boldsymbol{t}_i \boldsymbol{R}$. By construction, $\boldsymbol{h}$ and each $\boldsymbol{x}_i$ has an angle of exactly θ. We sample $n_{\text{train}} \in \{150, 500, 2000\}$ draws of the training examples of $\boldsymbol{t}_i$ and set $T_n := \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ as the training set of n_{train} examples with exactly $n_{\text{train}}/2$ examples with angular separation θ and $n_{\text{train}}/2$ examples with angular separation $-\theta$ by multiplying half the vectors by -1. We label $\boldsymbol{y}_i = 1$ if the corresponding $\boldsymbol{x}_i$ has an angular separation θ and $\boldsymbol{y}_i = -1$ if the angular separation is $-\theta$. As noted in Durrant and Kabán (2013), θ can be interpreted as the difficulty of the classification problem with a value of θ that is closer to 90° representing a more difficult problem with a "smaller margin" separating the two classes, so this process generates a sequence of increasingly difficult problems. Furthermore $\boldsymbol{h}$ is the Bayes' optimal classifier which separates the two classes perfectly with

[2]Source code for our experiments is available at `http://www.github.com/nlim-uow/majorityvote/`

maximum margin w.r.t the generating process, and it is not aligned with any coordinate axis with probability 1.

Using the same data generation scheme, we also generate an additional $n_{\text{val}} = 1000$ and $n_{\text{test}} = 1000$ model validation and test examples, with the corresponding class labels $\{1, -1\}$, divided evenly within all datasets. Then $\boldsymbol{X}_{\text{train}}, \boldsymbol{X}_{\text{val}}$ and $\boldsymbol{X}_{\text{test}}$ are the data matrices representing the training, holdout, and test data respectively. For each set of experimental conditions, we learned $N = 250$ base classifiers, using the following diversity generation schemes and learning algorithms, for each projection dimension $k \in 2, 5, 10, 20$:

- Linear Discriminant using random subspace projection, with the routine provided by Matlab Central (Dwinnell, 2010)
- Linear Discriminant using Orthonormalized Gaussian random projections
- L2 Regularized Linear Kernel Support Vector Machine with random subspace projection and the liblinear routine of Fan et al. (2008).
- L2 Regularized Linear Kernel Support Vector Machine with orthonormalized Gaussian random projection and the liblinear routine of Fan et al. (2008).
- Random Forest with Bag Size = $1/3\ n_{\text{train}}$ with the routine provided by Matlab.

We constructed an ensemble of $N = 250$ classifiers using these base classifiers, and measured the empirical majority vote accuracy on the test data $\boldsymbol{X}_{\text{test}}$ using the zero-one loss. In our experiments each ensemble member is learned from (a particular RS or RP projection of) the training data $\boldsymbol{X}_{\text{train}}$. For each set of experimental conditions, we repeated this process 30 times. We extracted the Sneath and Sokal (1963) diversity measure ρ and the average individual classifier accuracy p for the Polya-Eggenberger model using the empirical results from the experiments in section 4.1. For comparison, we also extracted the Q-statistics, the vote cosine similarity ('Vote Correlation') score, and the Jaccard similarity index to compare against the ρ diversity measure, as well as the simple binomial model for ensemble accuracy ($\psi = 0$). In each case the parameters of the corresponding PE models are estimated by applying the ensemble learned from the training data $\boldsymbol{X}_{\text{train}}$ on the validation data $\boldsymbol{X}_{\text{val}}$. We then calculated the CDFs for the Polya-Eggenberger models corresponding to these different diversity measures and we compare the predicted ensemble accuracy modelled by the Polya-Eggenberger distributions to the empirical majority vote ensemble accuracy on the test data $\boldsymbol{X}_{\text{test}}$ averaged over 30 runs.

Figure 1 shows the predicted ensemble accuracy modelled by the Polya-Eggenberger Distribution with the different determination of ψ against the empirical majority voting ensemble accuracy averaged over 30 runs on our synthetic data.

Tables 3 shows the comparison between the average majority vote ensemble classifier accuracy for ensemble size $N = 50$, $N = 100$ and $N = 250$ against the values predicted by the Polya-Eggenberger Model using the Sneath and Sokal (1963) diversity measure as the estimate for ψ.

4.2 Experiments using Real Datasets

We used six real-world datasets, three taken from the 2003 NIPS feature selection challenge, namely GISETTE, ARCENE and DEXTER (Guyon et al., 2004), and data from MNIST handwriting recognition challenge, IMDB sentiment analysis, and GTZAN music/speech recognition.

Using the same base classifier families as for our synthetic test cases, we learnt $N = 100$ classifiers for each of the base classifiers.

For the MNIST handwriting challenge, the task is to discriminate between digits '7' and '9' from low-resolution images. In the IMDB dataset, the task is to discriminate between positive and negative reviews using bags-of-words data from individual reviews. For the IMDB data we did not use random projection as a diversity generator as the time and space complexity for the random preprocessing was exorbitant. Finally, for the GTZAN music/speech classification task, we segmented the audio samples into 1-second segments and trained on the odd-second segments and tested on the even-second segments. The characteristics of the datasets and the number of projection dimensions used is summarized in Table 2.

For the real world data we removed the features that had zero variance in the training set, but to avoid possible confounds we carried out no other feature engineering. Apart from the smaller ensemble sizes ($N = 100$), also we made a 50:50 random split of the test data and fitted the P-E parameters on one half for these experiments. Evaluation of the model accuracy is on the remaining held-out test data, similarly to the experiments on synthetic data. Figure 3 shows the predicted ensemble accuracy modelled by the Polya-Eggenberger Distribution for the different datasets, with the different estimators of ψ and the binomial model ($\psi = 0$), against the empirical majority voting ensemble accuracy averaged over 100 runs. Results presented here are representative of typical outcomes, but for more complete findings for all experiments please see the supplementary material.

Table 4 shows the largest difference between the aver-

Table 2: Characteristics of the non-synthetic datasets.

File	Type	Train Set	Test Set	# feat	Proj. Dim
ARCENE	Dense	100	100	10000	2,10,20
DEXTER	Sparse Integer	300	300	20000	20,100, 200
GISETTE	Dense	6000	1000	5000	2,10,20
MNIST	Dense	60000	10000	784	2,10,20
IMDB	Sparse Integer	25000	25000	89527	200,1000, 2000
GTZAN	Dense	1920	1920	22050	20,100, 200

age majority vote ensemble classifier accuracy against the values predicted by the Polya-Eggenberger Model using the Sneath and Sokal (1963) diversity measure as the estimate for ψ.

4.3 Comments on Experiments

Table 3: Comparison between the empirical majority vote ensemble accuracy (left) and our model (right) for the synthetic data of Section 4.1 with many irrelevant features. Observe that the actual values were within 1% of the model predictions.

		k=2		
theta	n	N=50	N=100	N=150
80	150	66.0 / 66.6	71.4 / 72.3	75.0 / 76.0
	500	66.7 / 66.9	73.4 / 72.7	76.7 / 76.4
	2000	67.1 / 66.9	73.6 / 72.6	76.5 / 76.4
85	150	58.6 / 58.6	61.6 / 61.8	64.3 / 64.0
	500	59.1 / 59.3	61.8 / 62.7	64.2 / 65.0
	2000	60.0 / 59.9	63.5 / 63.5	65.8 / 66.0
87.5	150	53.6 / 53.6	55.0 / 55.0	56.3 / 55.9
	500	53.9 / 54.0	55.6 / 55.5	56.9 / 56.6
	2000	54.3 / 54.4	56.2 / 56.1	57.4 / 57.3

We can see that the Polya-Eggenberger model using the Sneath and Sokal (1963) diversity measure provides a very good estimate for the average majority vote ensemble classifier accuracy. The difference between the majority vote accuracy is visually almost imperceptible as seen in Figure 1. The largest differences in performance from our model occur for the smallest values of the projection dimension k, with the ensemble member size approximately $N/2 = 125$, in which case the absolute empirical majority vote accuracy differs by less than 2% compared to the Polya-Eggenberger model. This is despite the fact that the individual classifiers do not have identical accuracies, with the standard deviation of the accuracy of the ensemble members $\approx 1.2\%$, indicating that violating this assumption is not fatal to good estimation of the ensemble accuracy. Table 3 summarizes differences between the empirical and the modelled accuracy for different experimental conditions for the synthetic data with the smallest $k = 2$.

We also see that our model gave a very good approximation to the majority vote ensemble accuracy on the real world datasets; with difference between the estimates and the empirical values within 3%. Here the gap between the different estimators is much more clear-cut. We have not run statistical tests of hypotheses between the different estimators – it is clear from the graphics that using the ρ diversity measure generally outperforms the other alternatives considered here.

Figure 3 shows a representative selection of comparisons between the empirical accuracy and the modelled accuracy for different classifiers on the real world datasets, while Table 4 summarizes the absolute difference between the empirical accuracy and our model. For the sake of full disclosure we note that for cases when k is set very small compared to d – for example smaller than $\log d$ – our model can fail to give an accurate approximation of ensemble performance, and exhibits non-monotonic behaviour. These figures are omitted here due to space constraints – A complete set of results and figures is available in the supplementary material.

5 Discussion and applications of our model

Our experiments support the view that we can use the Polya-Eggenberger distribution, with the dispersion parameter estimated using Sneath and Sokal (1963) to model the accuracy of a majority vote ensemble classifier faithfully. A limitation of this model is that it assumes equal accuracies across ensemble members and this is not true in general. However, as our empirical results indicate, this assumption is not crucial to good performance of the model. Moreover the Polya-Eggenberger model gives a very good estimate of the average accuracy of the majority vote ensemble across a very wide range of ensemble member sizes, and independently of those particular aspects of the problem that are not related to classifier diversity, such as the original number of predictor variables, d, the hypothesis class of functions from which the model is selected, and even the sample size, n. We note that these factors must of course play a part in the accuracy of the base classifiers, and therefore they have an indirect effect on the parameters of our model. Nonetheless it is surprising, to us, just how well we can model ensemble accuracy without considering all of those important ingredients.

Consider two classification ensembles with identical average classifier performances $p > 0.5$, the first with classifier correlation r and ensemble member size N,

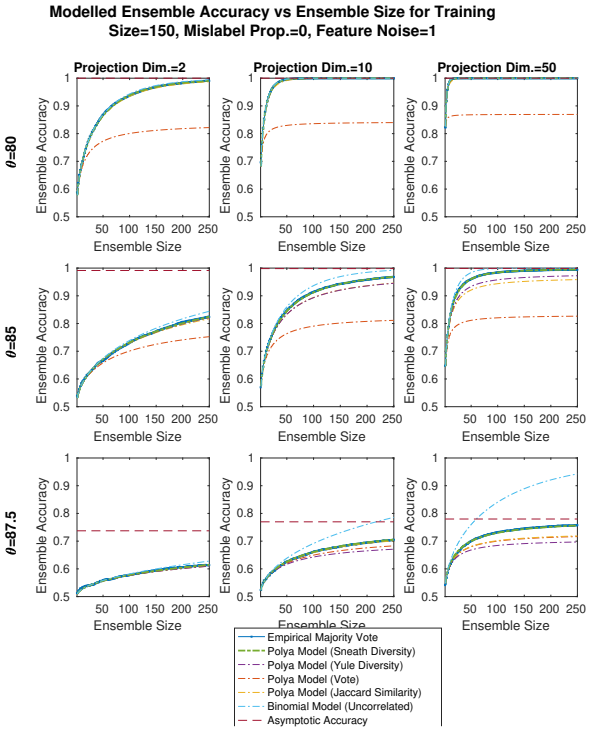

Figure 1: Majority vote ensemble accuracy on synthetic data as modelled by a Polya-Eggenberger distribution vs ensemble size for a small sample size setting $n = 150$ and dimensionality $d = 1000$. The solid line is the average majority vote accuracy measured empirically. Dashed lines are the accuracies modelled using different diversity measures. The Sneath diversity model most accurately estimates the empirical majority vote accuracy.

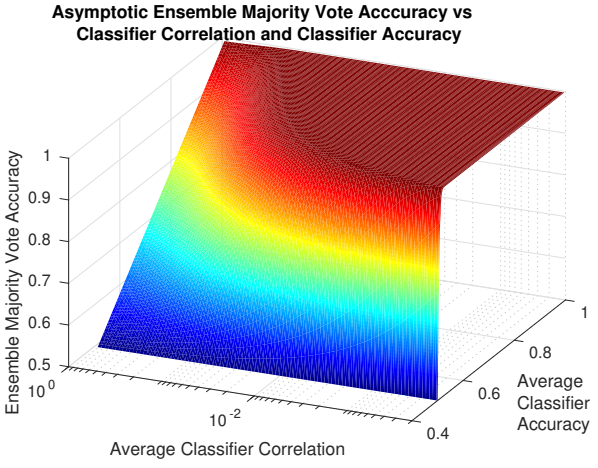

Figure 2: Surface plot for the asymptotic accuracy of a majority vote ensemble with $N \to \infty$

and the second with classifier correlation r' and ensemble member size N' : If $r + \frac{1-r}{N} < r' + \frac{1-r'}{N'}$ then

Table 4: Maximum absolute residual between the empirical majority vote ensemble accuracy and our model on real world datasets. The results marked with (*) are ensembles with non-monotonic accuracies. Observe that on these data our model estimates were always within 5% of the empirical accuracy.

		Maximum Residual		
Dataset	Base Classifier	k=2	k=10	k=20
ARCENE	LDA-RS	4.55*	2.28	1.31
	LDA-RP	3.96	1.83	2.00
	SVM-RS	4.56*	2.18	1.46
	SVM-RP	3.08	3.00	1.56
	RF	1.57	1.31	1.96
GISETTE	LDA-RS	1.97	1.35	0.67
	LDA-RP	0.7	0.54	0.75
	SVM-RS	1.36	2.32	1.24
	SVM-RP	0.64	0.66	0.66
	RF	1.93	1.50	1.06
MNIST	LDA-RS	4.36*	1.72	2.38
	LDA-RP	1.04	1.35	0.51
	SVM-RS	4.46*	1.81	1.1
	SVM-RP	1.43	1.29	0.5
	RF	2.04	0.79	0.49
Projection Dimension		k=20	k=100	k=200
DEXTER	LDA-RS	3.76*	1.20	1.19
	LDA-RP	1.34	1.04	0.95
	SVM-RS	1.85	1.47	0.87
	SVM-RP	2.00	1.58	0.77
	RF	3.1	1.26	1.81
GTZAN	LDA-RS	1.59	1.07	0.90
	LDA-RP	1.26	1.01	0.82
	SVM-RS	1.53	1.04	0.65
	SVM-RP	1.24	0.83	0.74
	RF	0.65	1.10	1.24
Projection Dimension		k=200	k=1000	k=2000
IMDB	SVM-RS	1.00	0.37	0.41
	RF	2.74	0.75	0.86

the ensemble on the LHS will typically have the better majority vote accuracy since the variance of the errors is smaller than that of the ensemble on RHS. Thus we have a simple criterion for model selection that does not require any nested (or indeed any other) structure to make sense of.

A further practical question of interest is whether it is better to have fewer negatively correlated classifiers (generated via careful selection), or (infinitely) many correlated classifiers. Here, one can verify that if we can generate $N > \frac{1-r}{r'-r}$, then fewer negatively correlated classifiers are better.

Since we can empirically predict the accuracy of different majority-voting ensembles precisely with our model we can also predict the effects of greedy - vs. non-greedy pruning of ensemble members by modelling the effect (on ensemble accuracy) of removing different ensemble members.

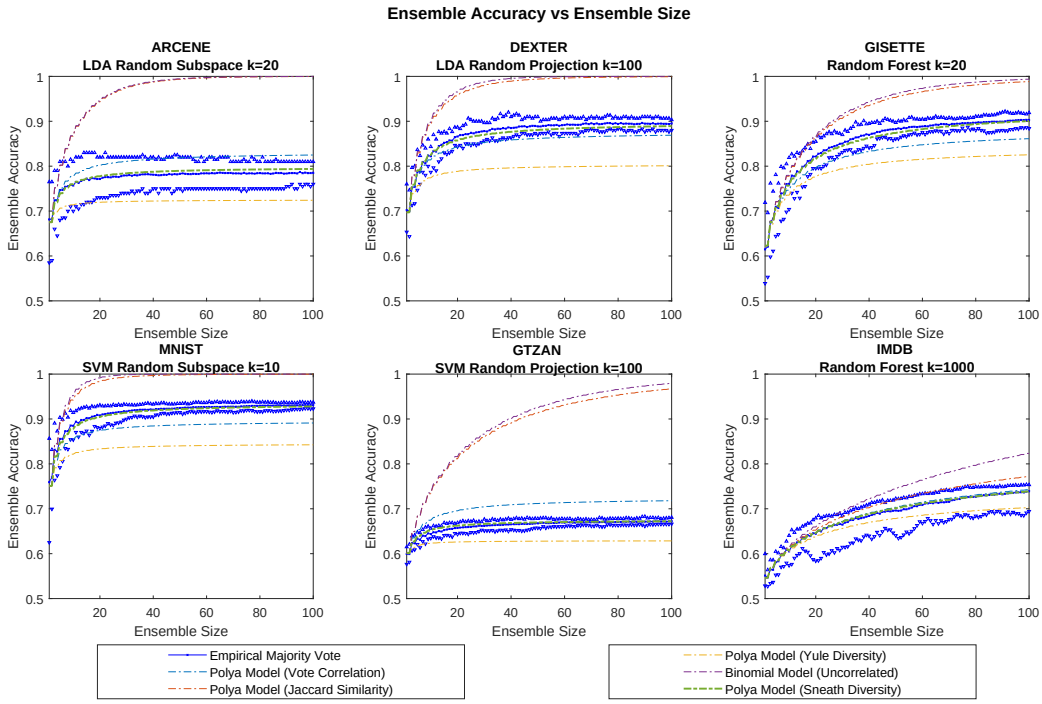

Figure 3: Average Majority vote ensemble accuracy on real-world datasets as modelled by a Polya-Eggenberger distribution vs ensemble member size for various datasets on different base classifiers. The upper and lower ticks are bootstrap confidence intervals for the upper-95% and lower-5% empirical accuracy. Observe that our model predicts the average ensemble accuracy almost exactly.

Now we consider the behaviour of a large ensemble under our model for scenarios when the (average) correlation of the classifier errors $\hat{r} = 0$, $\hat{r} < 0$ or $\hat{r} > 0$. When $\hat{r} = 0$ and average accuracy $\overline{p}$ is greater than 0.5 the P-E distribution has the same form as a binomial distribution, and CJT then implies the majority vote ensemble accuracy will increase as the ensemble size N tends to infinity. However we see from the model that for $\hat{r} \leq 0$ accuracy of the ensemble should improve, consistent with the findings of Kuncheva et al. (2000). However it is not possible for **all** classifiers to be both accurate **and** have negatively correlated errors in a majority vote, there is a strict lower bound on ψ of $-1/N$ and this tends to zero from below as $N \to \infty$, so $\psi \nearrow 0$ approaching a binomial model. One can think of this roughly as a consequence of the pigeonhole principle.

Finally, if $\hat{r} > 0$, observe that the P-E distribution is equivalent to a beta-binomial distribution with parameters $\alpha = \overline{p}\frac{1-r}{r}$, $\beta = (1-\overline{p})\frac{1-r}{r}$ and N. The limiting distribution as the number of ensemble members $N \to \infty$ is the beta distribution with parameters α and β. Using the CDF for the beta distribution we can explore the asymptotic behaviour of a large ensemble with $\hat{r} > 0$. Figure 2 illustrates the CDF for various values of α and β. We see that when $\overline{p}$ is close to 0.5, the size of ensemble N required to approach the asymptotic behaviour can be very large. As a general rule of thumb it seems that the ensemble size should be at least $N \in O(\frac{1}{(p-0.5)^2})$ before the majority vote ensemble classifier accuracy tends to the estimated asymptotic accuracy given by the corresponding beta model. This knowledge may have practical value for ensemble classification implementations where space considerations (on a smart device, for example) limit the ensemble size.

6 Conclusions and Future Work

We showed the accuracy of a majority-voting ensemble can be accurately modelled using a Polya-Eggenberger model using the Sneath and Sokal (1963) diversity measure (ρ). We discussed some implications of our model and verified it empirically on synthetic and real data. Surprisingly the accuracy of our model seems robust to properties that tend to make accuracy estimates for single models less robust, such as high dimensionality or small sample size. We note these affect the accuracy of individual ensemble members, the average of which is a parameter in the P-E model we propose. Modelling the accuracy of weighted majority-voting ensembles remains a challenging open problem.

References

Berg, S. (1993). Condorcet's jury theorem, dependency among jurors. *Social Choice and Welfare*, 10(1):87–95.

Brown, G., Wyatt, J., Harris, R., and Yao, X. (2005). Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20.

Condorcet, M. d. (1785). *Essay on the Application of Mathematics to the Theory of Decision-Making*. De L'Imprimerie Royale, Paris, France.

Durrant, R. J. (2013). *Learning in high dimensions with projected linear discriminants*. PhD thesis, University of Birmingham.

Durrant, R. J. and Kabán, A. (2013). Sharp Generalization Error Bounds for Randomly-projected Classifiers. In *Proceedings of the 30th International Conference on Machine Learning, (ICML)*, volume 28, pages 693–701, Atlanta, GA. JMLR.org.

Dwinnell, W. (2010). LDA: Linear discriminant analysis. Retrieved from: `https://www.mathworks.com/matlabcentral/fileexchange/29673-lda-linear-discriminant-analysis`. [Online; accessed 4-Dec-2018].

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Feller, W. (2008). *An introduction to probability theory and its applications*, volume 2. John Wiley & Sons, Hoboken, NJ.

Guyon, I., Gunn, S., Ben-Hur, A., and Dror, G. (2004). Result analysis of the nips 2003 feature selection challenge. In Thrun, S., Saul, L. K., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16, (NIPS)*, pages 545–552, Vancouver and Whistler, Canada. MIT Press.

Kuncheva, L. I. and Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207.

Kuncheva, L. I., Whitaker, C. J., Shipp, C. A., and Duin, R. P. (2003). Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis & Applications*, 6(1):22–31.

Kuncheva, L. I., Whitaker, C. J., Shipp, C. A., and Duin, R. P. W. (2000). Is independence good for combining classifiers? In *15th International Conference on Pattern Recognition, (ICPR)*, volume 2, pages 168–171, Barcelona, Spain. IEEE Computer Society.

Ladha, K. K. (1995). Information pooling through majority-rule voting: Condorcet's jury theorem with correlated votes. *Journal of Economic Behavior & Organization*, 26(3):353–372.

Lam, L. and Suen, S. (1997). Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics*, 27(5):553–568.

Malmasi, S. and Dras, M. (2015). Language identification using classifier ensembles. In Nakov, P., Zampieri, M., Osenova, P., Tan, L., Vertan, C., Ljubešić, N., and Tiedemann, J., editors, *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 35–43, Hissar, Bulgaria. Association for Computational Linguistics.

Nogueira, S. and Brown, G. (2015). Measuring the stability of feature selection with applications to ensemble methods. In Schwenker, F., Roli, F., and Kittler, J., editors, *International Workshop on Multiple Classifier Systems*, volume 9132 of *Lecture Notes in Computer Science*, pages 135–146, Günzburg, Germany. Springer.

Sen, K. and Mishra, A. (1996). A generalised Polya-Eggenberger model generating various discrete probability distributions. *Sankhyā: The Indian Journal of Statistics, Series A*, 58:243–251.

Sneath, P. H. and Sokal, R. R. (1963). Principles of numerical taxonomy. *Taxon*, 12(5):190–199.

Whalen, S. and Pandey, G. (2013). A comparative analysis of ensemble classifiers: Case studies in genomics. In Xiong, H., Karypis, G., Thuraisingham, B. M., Cook, D. J., and Wu, X., editors, *IEEE 13th International Conference on Data Mining*, pages 807–816, Dallas, TX. IEEE Computer Society.

Whitaker, C. and Kuncheva, L. (2003). Examining the relationship between majority vote accuracy and diversity in bagging and boosting. Technical report, School of Informatics, University of Wales, Bangor.

Yang, L. (2011). Classifiers selection for ensemble learning based on accuracy and diversity. *Procedia Engineering*, 15:4266–4270.

Yule, G. U. (1900). On the association of attributes in statistics: with illustrations from the material of the childhood society. *Philosophical Transactions of the Royal Society of London*, 194:257–319.