Hybrid Expert Ensembles for Identifying Unreliable Data in Citizen Science

Pieter Wessels[1], Nick Moran[2], Ali Johnston[2,3,4], Wenjia Wang[1]*

1: School of Computing Sciences, University of East Anglia, Norwich, UK.
2: British Trust for Ornithology(BTO), Thetford, UK.
3: Conservation Science Group, Department of Zoology, University of Cambridge, UK.
4: Cornell Lab of Ornithology, Cornell University, USA

Abstract

Citizen science utilises public resources for scientific research. BirdTrack is such a project established in 2004 by the British Trust for Ornithology (BTO) for the public to log their bird observations through its web or mobile applications. It has accumulated over 40 million observations. However, the veracity of these observations needs to be checked and the current process involves time-consuming interventions by human experts. This research therefore aims to develop a more efficient system to automatically identify unreliable observations from large volume of records.

This paper presents a novel approach – a Hybrid Expert Ensemble System (HEES) that combines an Expert System (ES) and machine induced models to perform the intended task. The ES is built based on human expertise and used as a base member of the ensemble. Other members are decision trees induced from county-based data. The HEES uses accuracy and diversity as criteria to select its members with an aim of improving its accuracy and reliability.

The experiments were carried out using the county-based data and the results indicate that (1) the performance of the expert system is reasonable for some counties but varied considerably on others. (2) An HEES is more accurate and reliable than the Expert System and also other individual models, with Sensitivity of 85% for correctly identifying unreliable observations and Specificity of 99% for reliable observations. These results demonstrated that the proposed approach has the ability to be an alternative or additional means to validate the observations in a timely and cost-effective manner and also has a potential to be applied in other citizen science projects where the huge amount of data needs to be checked effectively and efficiently.

Keywords: Citizen science, Ensemble, Expert system, BirdTrack, Classification, Diversity.

1. Introduction

Citizen science that engages the public to make their contributions to a designated topic has been rapidly increasing its popularity during recent decades and plays an important role in some research areas (Bonney et al., 2009; Wiggins et al., 2011; Wiggins and He, 2016), particularly on observations of the natural world such as wildlife sightings.

The British Trust for Ornithology (BTO) is a charitable organisation that monitors bird populations in the UK. On behalf of Royal Society for the Protection of Birds (RSPB), BirdWatch Ireland, Scottish Ornithologists' Club and Welsh Ornithological Society, the BTO created a citizen science project called BirdTrack to encourage members of the public to submit their observations of birds, including Time, Location, Observer, Observed bird species, and Number of individual birds seen, etc. To date over 40 millions of such observations have been collected.

Naturally, observers vary in experience, so the submitted observations need to be screened for anomalies before the data can be used for further analysis. The most common anomaly is that species are misidentified, particularly for rare species. Screening is done manually by a volunteer network of regional validators and is certainly time-consuming. Validators can use a set of locally-set filters on rarity, count and early/late dates for migratory species and more importantly their own local knowledge and experience to judge whether an observation is reliable or unreliable. Observations are labelled: "reliable", "unreliable" or "under query". Data are also checked at a national level but resources for doing this are limited. Sometimes unreliable records are not identified in the first instance, but are picked up by

[*] Corresponding author: wenjia.wang@uea.ac.uk

data consumers when they attempt to analyse the data and reported back to the data custodians.

Given the trend of increasing public participation, the rate of submitted observations will increase rapidly. Consequently, the burden of validating the observations will continue to increase and has the potential to overwhelm volunteer validators and ultimately compromise the quality of the data (Lewandowski and Specht, 2015; Lukyanenko et al., 2016; Bowser and Cooper, 2017). Therefore, there is an urgent need to consider alternatives such as artificial intelligence and machine-learning systems, to do the task in a more efficient and effective way. This motivates our research to develop an ensemble of expert systems and machine learning classifiers to identify unreliable observations in the BirdTrack data.

This research will also explore several issues that should be considered when building a hybrid expert ensemble system, including, specifically, how an expert system can be built based on human expertise and combined with machine induced classifiers, whether diversity among the members can be used to build more reliable and accurate ensembles, and the number of members needed for an ensemble. Our empirical results provide some useful guidelines for applying ensemble methods not only to this specific task but also to a wide range of applications in citizen science projects in general.

The rest of the paper is organised as follows. Section 2 reviews work related to the research, with a focus on methods for data validation on citizen science projects. Section 3 describes the proposed hybrid expert ensemble framework. Section 4 explains the data, preprocessing, partition and selection strategies. Section 5 presents the experiment design and results. Section 6 evaluates and discuses the work and results presented in the paper. Section 7 gives conclusions and suggestion for further work.

2. Related Work

The research on checking and validating the quality of citizen science data has been carried out almost at the same time as the citizen science was established simply because it was clear that the quality of data contributed by public varied significantly. Here we briefly review some important related work.

Bonney et al. (2009) presented a protocol for some citizen science projects, such as eBird – a continent-wide bird monitoring program developed and run by the Cornell Lab of Ornithology (CLO). Their protocol consists of up to nine components and they emphasised that "data quality is a critical issue for any citizen science project". Although they identified three measures that could be used to ensure that the collected data are as accurate as possible, there seemed no mechanism implemented at time of this study to validate the data entered into the system.

Wiggins et al. (2011) carried out a survey on the mechanisms used for data quality and validation in citizen science and found as many as 18 methods employed in various projects. Of them, expert review is the most common mechanism, employed in as many as 85% of about 50 surveyed citizen science projects. This is obviously expensive or at least labour-intensive if conducted by volunteers, and does not readily scale up for large scale citizen science projects. They pointed out that "one solution is applying data mining methods from computer science or collaboration with researchers in this area."

The CLO did a case study (Bonter and Cooper, 2012) on data validation for Project FeederWatch as they realised that "to become more widely accepted as valuable research tool, citizen science projects must find ways to ensure that data gathered by large numbers of people with varying levels of expertise are of consistently high quality." They designed a data validation system for this project, which consists of some automated filters to flag potential errors in bird observations for expert review. These filters were primarily built with some check-lists based on some simple statistical counts devised by the experienced researchers. This semi-automated system was tested on about 3.9 million submitted observations, 1.3% of them were flagged out for review and 97.7% were approved. However, there was no further breakdown for positive and negative cases because the data are obviously very unbalanced (much more valid cases than invalid, i.e. positive cases). Nevertheless, this case study demonstrated the feasibility and potential for using automated or semi-automated methods for data validation and hence this idea inspired us to develop an expert system to become a core member of our hybrid ensemble system.

Another study (Wiggins and He, 2016) adopted a mixed approach that involved relevant community members participating in data validation. They carried out a case study called iNaturalist. A sequential approach mixed some methods was used for validating the data through interactions between community members and the researchers/systems at different stages via various devices (e.g. PC and Mobile etc.). They found this approach was quite effective but its success was governed by several factors including the experience of participants and the devices they used. This approach still relies heavily on human participation and thus suffers from the very same issue that exists in any citizen science project, i.e. the variable experience in humans, in the first place. However, it should be noted that this mixed approach, technically effective, bears a high degree of similarity to our ensemble approach presented in this paper.

In summary, these related studies directly or indirectly pointed out that it would be more cost-effective to employ artificial intelligence and machine learning methods to build more accurate and reliable automated or semi-automated systems to do the job. This is the motivation of our research in developing a novel approach – a hybrid ensemble system that combines an expert system and machined induced classifiers to validate the bird observations more reliably and efficiently.

3. Hybrid Expert Ensemble Systems (HEES)

An ensemble is a machine learning paradigm that combines the output of multiple individual models by a decision fusion function with an aim of achieving more accurate and reliable solution for a given problem. Many ensemble methods (Dietterich, 2000; Wang et al., 2001; Anifowose et al., 2016) have been developed and applied to various problems (Perikos and Hatzilygeroudis, 2016; Kowalski et al., 2017; Liang et al., 2018). However, an ensemble does not necessarily perform better than individuals and a key factor that makes it successful is that its members must be diverse enough from each other(Yousefnezhad et al., 2016), i.e. having different strengths and weaknesses to avoid making the same mistake when working together, as obviously an ensemble built with identical members does not improve at all.

Many studies (Kuncheva and Whitaker, 2003; Wang, 2008; Richards and Wang, 2012; Rayana and Akoglu, 2016) have indicated that the models generated from different learning algorithms such as neural networks and decision trees, etc. are more likely to be more diverse than the ones generated from the same algorithms, but no study has yet proposed to include any expert systems that are built based on human knowledge, which is very useful and could be more diverse but difficult to be represented by machine models induced from available data. This is the main reason that this study proposes a novel ensemble framework that combines an expert system with several other machine generated classifiers to form a hybrid ensemble.

3.1. Hybrid Ensemble Systems

Notation: Let E be a hybrid ensemble; h_i a member model (classifier) in E; S the decision fusion function; N, the number of models in E; c_i, a candidate classifier in a pool C of machine generated classifiers, i.e. $c_i \in \{C\}$; $acc(c)$ and $acc(E)$ the accuracy of an individual classifier and an ensemble respectively; $D(E, h_i)$, a diversity measure between ensemble E and h_i.

A framework of the proposed hybrid expert ensemble system is depicted by Figure 1. As can be seen, it consists of an Expert System, h_1, and several machine learning induced classifiers $\{h_2, h_3, ..., h_n\}$, and their outputs are aggregated with a decision fusion function S.

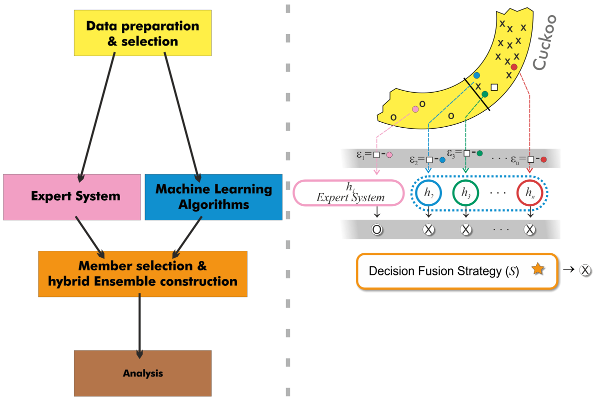
Data preparation & selection
Expert System
Machine Learning Algorithms
Member selection & hybrid Ensemble construction
Analysis
Cuckoo
ε1=□-●
ε2=□-●
ε3=□-●
εn=□-●
h1 Expert System
h2
h3
hn
Decision Fusion Strategy (S)

Figure 1: The framework of the proposed hybrid expert ensemble system shown on the left. The right figure illustrates the process for building a hybrid expert ensemble system, ideally with the ES and its complementary machine induced classifiers, $h_i, i = 2, ..., n$

Construction of the hybrid ensemble requires confirmation that the output produced by the Expert System is compatible with the output generated by the other classifiers and follows some strategies and rules to produce hybrid ensembles with various combinations based on the Expert System.

The main ideas and steps of a construction strategy are as follows with the assumptions that the expert system and a pool of candidate classifiers C have been built (details explained in the next section).

- Take the Expert System as the first member of the ensemble, h_1.
- Select the most accurate classifier from pool C of n machine generated candidate classifiers: $C = \{c_1, c_2, ..., c_i, ..., c_n\}$, as the second member, h_2, i.e. $h_2 = max\{acc(c_i)\}, \forall i = 1, ..., n$.
- Choose $c_i \in \{C \nsubseteq E\}$, to be $h_j, j = 3, ...,$ where: c_i has not already been included in the ensemble E, and is able to introduce a maximum diversity when added to the existing E, i.e. $D(E, h_j) = max\{D(E, c_i)\}$.

- Repeat the last step until the number of members in E reaches a pre-set size N.

The rules above attempt to control factors which influence the accuracy of an ensemble in the experiment and formalise the influence diversity has on the accuracy of the ensemble.

3.2. *Expert System and Machine Induced Classifiers*

The hybrid nature of the proposed ensemble system originates from the combination of expert system and machine induced classifiers. The former is intended to utilise the knowledge and experience of human experts, which is difficult for the latter to represent as machine induced classifiers can only learn from the data with a learning algorithm and the data usually does not contain human's experience or common sense. So, each of them is likely to capture different aspects of the underlying problem and hence, possibly, be more diverse from each other, and the combination of these two types of models in an ensemble provides a mechanism to utilise their own strengths to compensate each other's weaknesses, and then produce more reliable and accurate identification of unreliable observations. This section describes how the expert system and machine classifiers are built.

3.2.1. *Rule based Expert System*

Expert observers have accumulated abundant experience from birdwatching and some of them have been involved in manually checking and validating the submitted observations. Despite labour intensive and time-consuming, they have been doing a reasonably good job by labelling a considerable amount of unreliable records. It is thus reasonable to use their expertise in some ways in the proposed ensemble system, so, an Expert System was built by following process adapted from Ciarratano and Riley (2005):

- Conduct feasibility study: human experts and other resources are identified and the size and scale of the Expert System is considered.

- Rapid Prototyping: experts' knowledge was acquired and represented with a set of rules, and an initial expert system was designed and implemented with SWI-Prolog.

- Refine the system: The system was tested on the validation data and then refined accordingly.

Two available veteran birdwatchers were identified as experts and their experience, combined with information gleaned from bird identification books (Couzens and Nurney, 2013; Vinicombe et al., 2014) are extracted, abstracted and represented by some simple rules as illustrated in figure 2. It should be noted that this rule-based expert system was developed with a trade-off between accuracy, simplicity and efficiency, which is a common practice when implementing an expert system.

The expert system is tested in various Phases (details given in later sections) with the data from three counties and the results, given in Table 1, show that it achieved good performance on one county, with the sensitivity (defined in Section 3.4) between 67% and 81% on True Testing data (explained in Section 4.2) even higher on Validation data, but poor on the other two. These poor results, however, are not surprising because the geographical distribution of bird species is often characterised with locality, and the experience of human experts could also be limited to their local areas, and the rules extracted from their knowledge are limited accordingly as well.

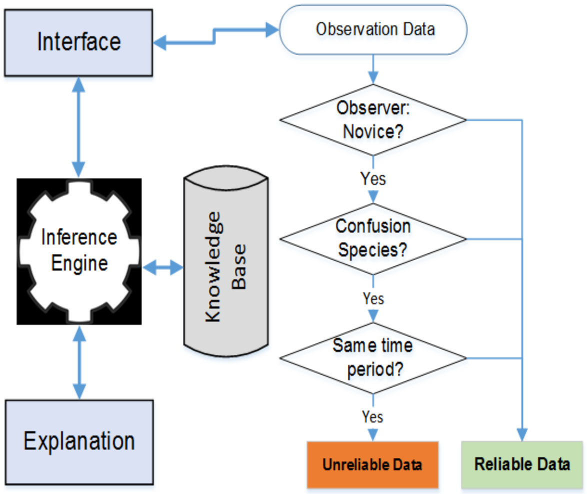

Figure 2: The Expert System and an example of the rules built within the expert system. An example of the explanation given by the ES on the novice is that who has submitted less than about 100 complete observations.

There is certainly room for improving the expert system in a variety of ways, such as adding more detailed rules, and/or devising sub-expert systems for each county or region, etc. But as this study is more about demonstration of the concept, on balance, it is considered to be acceptable as a core member of the hybrid ensemble system for two reasons: (1) After all it represents human expertise to some extent, which could not be directly learned through machine learning, and (2) it can thus provide some complementary contribution to

Table 1: Sensitivity of the expert system per core, per phase on three representative counties: GBSTA, GBWAW and GBWMI.

GBSTA		GBWAW		GBWMI		Phase
S	C-fold	S	C-fold	S	C-fold	
0.52	0.34	0.33	0.55	0.62	0.38	Phase 1
Validation		Validation		Validation		
	0.30		0.86		0.28	Phase 2
	0.30		0.64		0.26	Phase 3
	0.34		0.83		0.28	Phase 4
True test		True test		True-test		
	0.32		0.67		0.17	Phase 2
	0.33		0.76		0.13	Phase 3
	0.26		0.81		0.22	Phase 4

the hybrid ensemble.

3.2.2. Generating Machine Learning Classifiers

In principle, any machine learning algorithms can be used to generate classifiers as the candidates for being considered as the members of a hybrid ensemble. Therefore, at the early stage of this study, a number of different algorithms were tried, including k-nearest neighbourhood (kNN) method and logistic regression. But due to the high dimensionality and quantity of the data, they failed to produce viable results, even using a high performance computing cluster. As a result a decision tree induction algorithm implemented in R, equivalent to C5 (Kuhn et al., 2015), has been used primarily because of its reasonably high accuracy, efficiency and more importantly transparency.

However, due to limits in its implementation in terms of handling the types of attribute and missing values, the continuous attributes needed to be discretized before ingestion into the program.

When inducing decision trees, four different data preparations produced in experiment Phases 1 and 2 were applied to each of these subsets: 3 cores, 2 segments each (S & C-fold), 2 data representations with discrete and continuous features to test which sets of the features are more relevant and useful (Aldehim and Wang, 2017; Cervantes et al., 2018), so several hundreds of classifiers were induced as the candidates for building hybrid ensembles. The details of data preparation and partition for training, validation and testing are described in Section 4.2.

3.3. Factors need to be considered when building an ensemble

As the ultimate objective of using an ensemble philosophy is to improve the accuracy of solution for a given problem, it is then important to know what internal factors influence the accuracy in order to build more accurate ensembles.

This has been an active area of research (Kuncheva and Whitaker, 2003; Wang, 2008; Richards and Wang, 2012) and different conclusions were produced. However, it is generally viewed that four factors (Wang, 2008) need to be considered when building an ensemble E: diversity D among the member modules in E, the accuracy of individual member models h_i, the size of E and the decision fusion strategy (S).

A function $f()$ relating accuracy of E and these four factors can be conceptually represented as:

$$acc(E) = f(acc(h_i)\{\forall i = 1 \text{ to } N\}, D(E), S, N) \qquad (1)$$

Where $acc()$ = accuracy, N = the number of member models h in E; $D(E)$, the diversity among the members $h \in E$; S, decision fusion method (Wang, 2008). f is a non-deterministic function that varies in accordance with the decision fusion function S and the other three factors: $acc(), D$ and N. So the relationship appears to be complex, non-linear and variable.

3.3.1. Diversity D

Diversity has been perceived to be a key issue affecting the accuracy of E. Kuncheva and Whitaker (2003) probed 10 different definitions for “diversity” and found that most of the existing diversity definitions are not effective, except the Coincident Failure Diversity(CFD) (Partridge and Krzanowski, 1997), consistent with the experimental evidence. This is the reason that the CFD is used in this study.

The CFD measures the probability that all models in an ensemble fail coincidentally on the test data and is defined as:

$$CFD = \begin{cases} \frac{1}{1-p_0} \sum_{k}^{N} \frac{N-k}{N-1} & \text{if } p_0 < 1 \\ 0 & \text{if } p_0 = 1 \end{cases} \qquad (2)$$

Where p_k is the probability that k members of E will make the wrong choice at the same time. p_0 is a special case where no member is wrong.

$CFD \in [0, 1]$, the larger, the more diverse the models in E are. When $CFD = 0$, it means there is no diversity among the models, i.e. all the models are identical, hence the ensemble has no gain on accuracy at all. When $CFD = 1$, it means that a maximum diversity is achieved among the models and the ensemble will produce perfect solutions on the test data.

3.3.2. Accuracy of individual classifiers, acc(h)

Another important factor that needs to be considered when building an ensemble is the accuracy of individual models h as in general, better member models lead to more accurate ensembles.

It is, however, difficult to determine how accurate of an individual model is good enough to be selected to be a member of an ensemble. This work (Wang, 2008) gave some guidelines based on an analysis of the relationship between ensemble's accuracy and individual member's accuracy. It defined a lower-bound accuracy for a classifier h:

$$acc(h)_{lb} = \lim_{N \to \infty} \frac{N+1}{KN} = \frac{1}{K} \tag{3}$$

Where K is the number of classes in a dataset being classified. It means that a classifier should achieve at least an accuracy equal to or higher than the lower-bound value $1/K$ to make some useful contribution to the accuracy of ensemble E. For example, for a binary classification problem, $K = 2$, then $acc_{lb} = 0.5$, which is also called the default accuracy of a classification problem.

So, in general, in order to improve the accuracy of E, the accuracy of all its members (hs) should be at least higher than the lower-bound.

However, even the accuracy of all members in E is higher than acc_{lb}, it is still possible that E may achieve a lower accuracy than the average of, or even the least accurate of its members (Wang, 2008) when the members work in a destructive manner. That is, they may cancel each out when they have a negative diversity. This possible phenomenon may occur in reality but has never been observed before and will hence be investigated in this study.

3.3.3. The size of ensemble, N

Conceptually speaking, the size matters, that means, the more members an ensemble has, the more reliable and accurate it could be. However, the problem is not as simple as that, because the accuracy of an ensemble is not solely determined by the size of it, simply because it involves other factors, $acc(h)$, diversity D, and decision making function S, as mentioned above. So in practice it is not clear how the size actually affects $acc(E)$ when coupled with the other factors, and what is the appropriate size for a given application problem. These issues were experimentally investigated in this study.

3.3.4. Decision fusion strategy, S

The decision fusion strategy, S is considered to be influential on the accuracy of an ensemble simply because it produces the final solution. The output of an ensemble can be determined when a specific fusion strategy S is chosen. Commonly used strategies include voting for classification and averaging for regression. But the relationship between the accuracy of ensemble $acc(E)$ and S is not deterministic because again $acc(E)$ cannot be determined by S alone, but also the three other above mentioned factors. This study employed the simple majority voting as it is commonly used as the decision fusion function for classification ensembles.

3.4. Metrics for performance evaluation

Determining appropriate metrics for evaluating the performance of classification is very important as they must be able to represent the classification accuracy on the target class in a quantitative manner, independent from the distribution of classes in a dataset. For these reasons, the following metrics are chosen to measure the accuracy of individual classifiers and ensembles in this research.

Sensitivity and Specificity are defined to measure the accuracy of a classifier in relation to positive class and negative class respectively. In this application, $Sensitivity$ represents the accuracy of identifying the positive records in the data, i.e. the unreliable records that were positively rejected by the human experts, and $Specificity$, the accuracy of identifying the records that were accepted by the human experts. They are defined as follows:

$$Sensitivity(h) = \frac{tp}{tp + fn} \tag{4}$$

$$Specificity(h) = \frac{tn}{tn + fp} \tag{5}$$

Where (with respect to the positive class):

- tp = true positives - number of observations in the positive class that are correctly classified as positive class by classifier h.
- tn = true negatives - number of observations not in the positive class, correctly identified.
- fp = false positives - number of negative observations that are falsely classified as belonging to the positive class.
- fn = false negatives - number of positive observations being incorrectly classified as negative class.

As the experiments were repeated several times in according to the experiment set-up and procedures, the average values of these measures are given in the result section.

4. Data, Preprocessing and Partition Strategies

The raw data need to be cleansed and pre-processed through several steps to make them ready for inducing classifiers to build ensembles. Cleansing may involve checking for invalid or contradictory records, invalid or irrelevant attributes, and/or missing values. Preprocessing involves feature transformation, discretization and normalisation etc., pending on the type of the machine learning algorithms employed in the study. Data selection determines how to select or partition the data into subsets for training, validation and test.

4.1. BirdTrack and Habitat Data

Two raw data sets were obtained: BirdTrack data and Habitat Data. BirdTrack data contain individual observations of bird species, time, location and observer details. Habitat data, from the Social and Environmental Economic Research (SEER) team, provides additional information on habitat types per square kilometre across the UK(Bateman et al., 2014).

The demographics of the BirdTrack data can be summarised as follows.

- 2,988,648 birdwatching lists, comprising
- 19,745,105 individual observations of 764 bird species/types [1]
- from 19,068 observers
- across 104,471 sites in 148 counties or special designated areas
- between 1 January 2007 and 30 April 2015.

The data were checked for quality and consistency, and it was found that some observations contain missing values or contradictions. For example, 95,338 observations missed observer identifier; 1,482 observations had no county details; and 24 species had the same ID, but multiple (similar) names. All those observations were deleted or manually corrected.

It was assumed throughout that the identification of the veracity of the observations was accurate – that is that the identification of unreliable bird observations was accurate and constant, as was the level of applied human expertise.

The habitat data provided combined habitat information for approximately 86% of the observations. The data were provided with the UK Grid Reference as a unique key. The distance to the nearest coastal or marine feature was calculated as the shortest Euclidean distance to any grid reference containing a marine or coastal feature. Because coverage was incomplete, it is possible that in some cases the distance to the nearest coastal or marine feature might not be accurate – that is, the algorithm will find the closest grid reference to those references it knows about, which might not be as close to the sea as a reference which is not included in the dataset.

The data sets were combined to become one dataset with 45 attributes. The detail of some important attributes in the dataset is given in Appendix A. Observations which were in query (i.e. had neither been accepted nor rejected by the BTO) were also excluded from the start

The data sets have been pre-processed through a series of steps, including discretization, feature transformation, normalisation etc. The details of all the features are given in Appendix A.

4.2. Partitioning data for training, validating and testing

After cleansing and pre-processing, the following rules are used to generate the data subsets (cores) for training and validating classifiers, and testing ensembles.

4.2.1. County-based data selection

As the distribution of bird species is commonly determined by location and time, different area represented by County can have their geographical characteristics, which could not be generalised, and hence may have their own localised bird species. Therefore it is reasonable to partition the whole data by county into county-based datasets. In doing so, it is hoped that the induced classifiers may learn more local knowledge and be more diverse from each other, which can result in more reliable and accurate ensembles. For this reason, three counties encoded as GBSTA, GBWAW and GBWMI were selected as the core datasets because their regional features were considered as more representative.

For each core dataset, a subset is firstly selected and kept aside as the ensemble testing data, called Ensemble Test. Then the remaining data, D_r, are further selected or sampled with two strategies: manual and random, to produce data subsets for training and testing classifiers.

4.2.2. Two strategies for selecting data: manual and random

As the experiments were designed and carried out in four phases, two strategies – manual and random, were

[1] This included rare and migrant species, sub-species, hybrids and in some cases generic “families” and catch-alls, e.g. “unidentified warbler”

used to select the data from D_r for training and validatingtesting, and their influences were investigated by their corresponding experiment phases. In Phase 1 the training data were manually selected by human expert from each core dataset. In the subsequent phases the data subsets were randomly selected for training and testing.

Data for each phase are stratified and randomly assigned to the subsets. The process of stratification ensures that each subset has a representative proportion of the minority class – this process is required because the distribution of the minority class is so small it cannot otherwise be guaranteed that a subset would contain examples from it.

*[*Manually selected data]Phase 1 - Manually selected data subsets: It was initially believed that human expertise should and could be utilised to select the data subsets that are more "representative" of the underlying problem and hence better models(classifiers) could be generated from the manually selected data, which may lead to build more accurate and reliable ensembles.

For each core dataset, three subsets: S-train, S-test and C-fold, were selected from D_r. This was done by select a subset of some years' data as training data S-train, and a subset of a different year's data as testing data, S-test. The C-fold subset was selected from D_r without any restriction and it was used for 10-fold cross-validation. Although the data contained in the C-fold and S-subsets may overlap, the Ensemble Test dataset is "independent" and unique.

The details of the three manually selected data subsets and their corresponding test dataset - $EnsembleTest$, are shown in Table 2.

Table 2: For Phase 1, the details of three manually selected data subsets(the numbers of instances are rounded to nearest thousand(k) for simplicity).

County	S-train	S-test	C-fold	Ensemble Test
GBSTA	138k	83k	428k	70k
GBWMI	72k	28k	258k	49k
GBWAW	73k	31k	263k	50k

*[*Randomly selected data]Phase 2 - Randomly selected data subsets:

In order to compare the results of ensembles trained by the manually selected data, a stratified random selection mechanism was employed to split the data into three non-overlapping subsets for training, validating and testing at a pre-fixed ratio as follows.

Select 70% of the available data as C-fold for training, 20% for validating the performance of the individual members and selecting the members for the ensemble, and keep the remaining 10% aside for testing.

The choice of these ratios (70/20/10) was primarily based on an empirical hypothesis that more training data are likely to produce better models. While this is not always true, as the focus of this research is to demonstrate the concept and not for trying to find the optimal training to test ratios, it was considered non-essential to try other partitioning ratios.

Details of the split are shown in table 3.

This procedure was repeated for the subsequent phases (3 and 4) with the same ratios so that the numbers of data instances in each subset are roughly the same.

Table 3: For Phase 2, the numbers of records in three randomly partitioned data subsets.

County	Train	Validation	Test
GBSTA	349k	100k	50k
GBWMI	215k	61k	31k
GBWAW	219k	63k	31k

5. Experiment Design and Results

5.1. Experiment Design

As mentioned earlier, the experiments were divided into four progressive phases. Each phase uses its corresponding data subsets, as described in the previous section, to generate the classifiers – decision trees – and then build ensembles with various strategies to test the performance of built hybrid ensembles and investigate the influence of diversity, accuracy of individual classifiers and the size of ensemble, on the accuracy of ensemble.

The generic procedure for building and testing hybrid expert ensembles is essentially the same for all the phases.

5.1.1. Generation of decision tree classifiers

Each S-train dataset is used for inducing a decision tree classifier and S-test dataset is used for validating. Their results were used to determine if a classifier is selected to add onto the ensemble.

The *K-fold cross validation* mechanism is applied to the C-fold data to generate decision trees. We set K to 10, so 10 classifiers are induced with 9 folds and validated with 1 remaining fold in one round-robin run for each data subset in each phase. Their validation performance was used to determine whether a generated model – decision tree, is good enough to be selected to add into an ensemble.

For each of the selected counties, around 60 decision tree classifiers were generated in each phase.

5.1.2. Strategies for building hybrid ensemble

As described in Section 3.1, after constructing the expert system and generating pools of decision tree classifiers for each county in each phase, hybrid ensembles can be built by using the expert system as a core member and then choosing the best classifier from a corresponding pool as another core member and some other classifiers based on the strategies designed for the purposes of investigations.

For investigating the relationship between accuracy and diversity, the following strategies are devised.

Strategy 1: Hybrid Ensembles of Fixed Size, or HEFS for short.

The size of hybrid expert ensembles in the experiments of Phase 1 is fixed in order to focus on the influence of accuracy of individual classifiers and diversity. The size is initially fixed to 3 (2+1) – the minimum number of members required for a valid ensemble.

After the first two core members have been established, the remaining candidates in the pool competes against each to become the third member of the ensemble. The third member can be chosen by applying some rules, such as comparing diversity measures with an aim to maximise the diversity of the ensemble.

The above strategy is applied when selecting more members to build larger ensembles. However, it should be noted that because the decision fusion strategy applied in this study is a simple majority voting, the number of members in an ensemble should be odd to avoid ties for binary class classification problems.

Strategy 2: Hybrid Ensemble of Growing Size, or Growing for short.

Another set of experiments are designed to investigate the relationship between accuracy and size of ensemble in a systematic manner.

In this set of experiments, the size of a hybrid ensemble is increased by one at a time with the most diverse candidate left in the pool until all the candidates have been included. The previous procedure can be used by a small modification on the last few steps. Therefore, the ensembles of fixed sizes then become only few special cases in this experiment design.

5.2. Experimental Results of Phase 1

Three sets of the results were produced from the experiments in Phase 1 and other phases as well and presented in this section. The first set is the sensitivity accuracy of individual classifiers on the testing data, $acc(c_i), c_i \in C$; the second set is the sensitivity accuracy of the ensembles $acc(E)$ and the average accuracy of the members in ensemble E; and the third set is the diversity in E measured by CFD, $div()$.

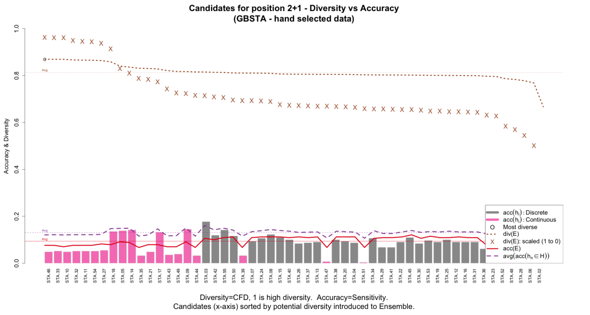

Figure 3: Phase 1: the diversity and accuracy results of 2+1 hybrid ensembles on hand-selected data subset GBSTA. The accuracy of individual classifiers $acc(h)$, accuracy of ensembles, $acc(E)$, average accuracy of the members in a hybrid ensemble $H = E$, $acc(h \in H)$, and the diversity $div(E)$ in E.

It should be noted that whilst all accuracies given are measured with sensitivity, the specificity is also measured but not presented here because it is always very high between 99% to 100% for almost all the cases.

5.2.1. HEFS 2+1: the battle for the third member

Figures 3 to 5 show the results produced in Phase 1 for the three subsets with discrete and continuous attributes respectively with many details.

The bar graph shows the Accuracy of each individual candidate, $acc(c_i)$ – colour coded to separate candidates trained on Discrete and Continuous data.

The third member in E is awarded to the candidate on the far left of the graph, as this is the candidate which is the most different from the existing ensemble. Notice that it may not be the most accurate $h_{candidate}$. Then, it is awarded to the next one in turn and so on until all the candidates have been used as the third member. In this way, ensembles as many as up to the number of the candidates have been built.

The accuracy of ensembles, $acc(E)$ is shown with a solid line. The average Accuracy of the individual members of the ensembles, $avg(acc(h_n \in E))$, is shown by the dashed line, assuming the candidate c_i has been accepted into the ensemble.

In addition the graphs also show diversity measure $div(E)$ on two scales. The first uses the same scale as the Accuracy, but as the changes in $div(E)$ are so slight, a second line is shown scaled so that the highest Diversity corresponds to 1 and the lowest to 0 - thus making it possible to see detail within the movement of $div(E)$).

As can be seen, the first two subsets(figures 3 and 4) are not good but the ensembles in the third subset (figure 5) are reasonably good.

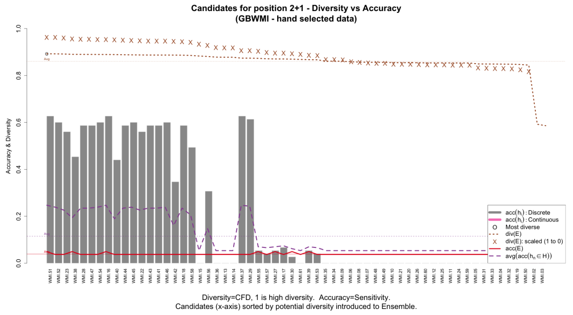

Figure 4: Phase 1: the diversity and accuracy results of 2+1 hybrid ensembles on hand-selected data subset GBWMI.

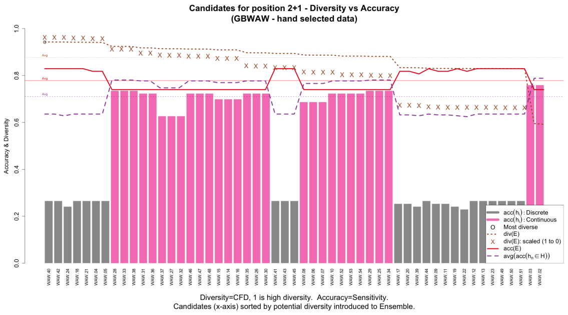

Figure 5: Phase 1: the diversity and accuracy results of 2+1 hybrid ensembles on hand-selected data subset GBWAW.

There are several notable characteristics in figure features 3 and 4):

- The machine induced decision tree classifiers are very bad, with the accuracy of only about 10 % on both discrete and continuous data sets. The accuracy of the ensembles is very bad as well.

- The accuracy of the ensembles, $avg(acc(E))$ lies below $avg(acc(h_n \in H = E))$, the average of the members. This means that the members of an ensemble make poor joint decisions when working together. Outside the ensemble, they are individually more likely to make the correct decision.

- Diversity $div(E)$ is very high – indicating that the decisions of the members differ significantly for any given case presented. Combined with the above phenomenon, $avg(acc(E)) < avg(acc(h_n \in H = E))$, it is clear that majority of the members are more likely to be on the wrong side, against one on the correct side for most of the test cases, hence the decision made by the ensemble is more likely to be wrong than the average performance of the members working individually.

The fact that $acc(c)$ is very low on the test data – even lower than a random guess– clearly indicates that these machine classifiers induced from the training data did not generalise well at all, and the main reason is that the hand selected training data is not representative on this dataset as intended.

Figure 4 shows similar characteristics to those discussed for dataset GBSTA, except there are some relatively good candidates – better than a random guess (shown by high grey bars).

It is noteworthy that none of the more accurate candidates have been trained on continuous dataset of the same data subset. The most diverse candidate (far left on the graph) also happens to be fairly accurate - it is added to E.

However, even with these relatively "good" candidates, the ensembles E performed still poorer than the average of the members in E. As the diversity $div(E)$ remains high, it shows again that the members work destructively to ignore the correct decisions made by minority members, so force the accuracy of the ensembles below that of the average of the individual members.

Figure 5 shows candidates competing for the third member in hybrid ensembles for data subset GBWAW. It is very clear that these ensembles are much better than the previous two situations as their accuracies are between 75% to 85%, and more importantly $acc(E)$ was sometimes about 10% more accurate than the average accuracy of the individual members, $avg(acc(h_n \in E))$, which is the evidence of the gain achieved by the ensembles utilizing the right diversity.

It is interesting to note that the candidate selected to join an ensemble E was not particularly accurate, but its inclusion into E resulted in an ensemble that performed better than the average of its members. On the other hand, the models trained on continuous data were more accurate than their counterparts but when they were added to hybrid ensembles E, $acc(E)$ dropped below $avg(acc(h_n \in E))$. This is probably because they are in this case relatively less diverse than those classifiers trained with discretized attributes.

Having produced ensembles of three members that can perform better than the average of the individual members, it would be interesting to explore what would happen if the ensemble size increases, e.g. to 5 and more.

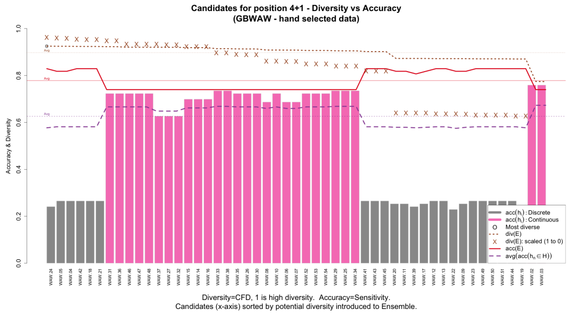

Figure 6: Phase 1: the diversity and accuracy results of 4+1 hybrid ensembles on hand-selected data subset GBWAW.

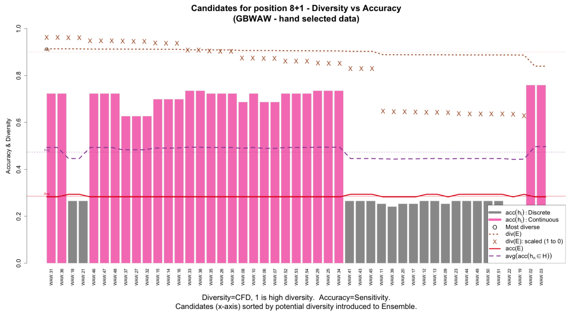

Figure 7: Phase 1: the diversity and accuracy results of 8+1 hybrid ensembles on hand-selected data subset GBWAW.

5.2.2. HEFS: 4+1 and 8+1 Ensembles

A hybrid ensemble E with 5 members is built with the following procedure. The expert system again is used as the first core member, then the most accurate three candidates in a pool are selected as the second, third and fourth members of ensemble E; the remaining candidates in the pool competes for the fifth member in E. So it is called as 4+1 ensemble.

Figure 6 shows the results of the 4+1 ensembles. It can be seen that (1) these ensembles consistently perform better (as $acc(E) > acc(h \in H = E)$) regardless of which candidate joins its ranks. (2) The more accurate individuals do not increase $acc(E)$ by as much as the less accurate candidates, just as observed in the previous case (2+1 ensembles on GBWAW data).

However, when the size of ensemble is increased to 8+1 using the same strategy as that of 2+1 and 4+1, the results of the hybrid ensembles, shown in figure 7, are very different from that of 4+1 ensembles, although the individual candidates seemed more or less the same.

The accuracy of these hybrid ensembles, $acc(E)$, is worse than the average of the members in E. In other words, although some of the individual members are quite accurate and also have relatively high diversity values, their addition to the ensembles make them worse collectively, which suggests again that this diversity is not helpful but actually harmful, therefore the diversity in this situation should be considered as negative.

So, these two cases of the enlarged hybrid ensembles demonstrated that it is not always true that the bigger an ensemble is, the better it can be. Then it is necessary to investigate how the size of ensemble may affect the performance of ensemble in a more systematic manner, when the size grows gradually from the minimum 3 to a possible maximum size equal to the size (around 60) of its corresponding candidate pool.

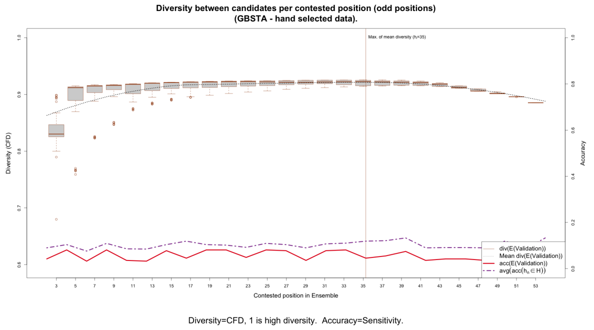

Figure 8: Phase 1: The diversity and accuracy results of growing ensembles hand-selected data. The relations between the accuracy and diversity in growing ensembles on GBSTA validation data. The Box plots show the diversity values(maximum, high quartile, median, lower quartile, minimum and outliers(circles) if any) of the ensembles for each size.

5.2.3. Growing ensembles

This section presents the results of growing hybrid ensembles to show how the accuracy of ensemble ($acc(E)$) may vary as the ensemble size (N) expands.

Figures 8 to 10 show the change in $acc(E)$ and $avg(acc(h_n \in H = E))$ as the number of members in a hybrid ensemble H expands from 3 to the maximum possible number - the size of the candidate pool, for the data of the three counties. The range of the $div(E)$ generated by all the candidates as they compete for each position is also shown. Note that the $div(E)$ rises until position 35 or 36 and then declines. Note also that the spread of $div(E)$ decreases - this is a natural consequence of removing the most diverse candidate from the pool of candidates each time.

Above all it is worth noting that $acc(E)$ never increases beyond that of $avg(acc(h_n \in E))$, which means that this set of hybrid ensembles is definitely destructive rather than constructive when making their final classification decisions. It is not a surprise considering that

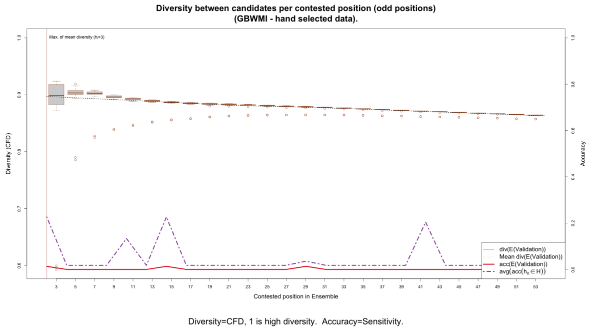

Figure 9: Phase 1: the diversity and accuracy results of growing ensembles on hand-selected GBWMI data. These ensembles are very poor, basically no use at all.

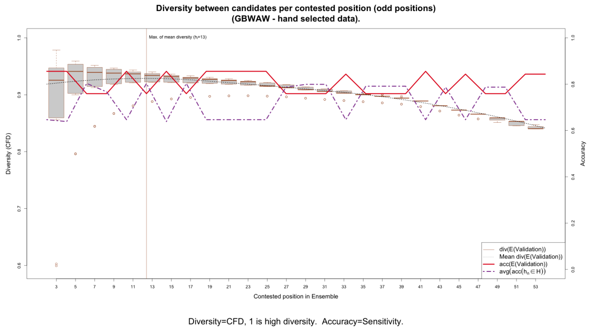

Figure 10: Phase 1: the diversity and accuracy results of growing ensembles on hand-selected GBWAW data.

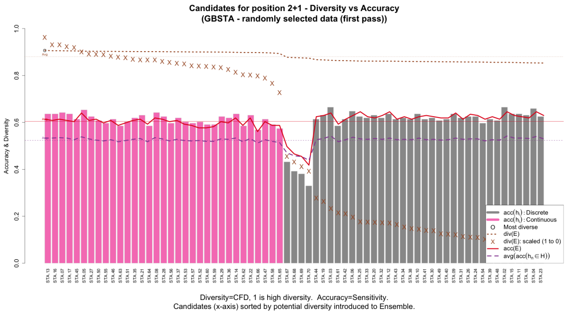

Figure 11: Phase 2: the diversity and accuracy results of 2+1 ensembles on randomly selected data GBSTA.

they are not “experts” at all - their decisions are worse than random (as shown in figure 3)

This set of the results demonstrates that, given a pool of very poor candidates, the accuracy of growing ensembles is not affected much by their size, nor the diversity, although it fluctuates a bit up and down.

The growing ensembles, shown in figure 9 built for data subset GBWMI do not differ much from those of GBSTA. All the ensembles E perform poorly again, even when fairly accurate members are added to them. Diversity peaks at position 3 and decreases as the committee grows.

The growing ensembles (as shown in figure 10) are in general much better as they increase their accuracy when the appropriate candidates are added in, but decrease when inappropriate candidates are selected to join the ensembles.

This inverse behaviour pattern, seen at cases 2+1, 4+1 and 8+1, is repeated many times over the whole length of candidate pool size dimension, and it is quite obvious that the size of an ensemble on average does not have much influence on the accuracy of ensembles built under the conditions as described in the earlier section, but the appropriateness of candidates does.

However, it is not clear what constitutes the appropriateness for a candidate as no study was found in the literature to define it explicitly. We think it should have something to do with a combination of suitable diversity between the members and accuracy of the candidates. The repeated patterns indicate that an appropriate candidate appears to be less accurate but more diverse in a certain direction, which has not been captured by the current diversity measure.

5.2.4. *Summary of the results of Phase 1*

Considering all the results produced in Phase 1 as a whole, one obvious conclusion can be self-evidently generalised, that is, the hand-selected data sets are not as representative as the human experts expected, or biased, because two sets of the ensembles built based on the hand selected data performed poorly, and only one set produced some reasonably good results. The possible reasons for this will be discussed later in Evaluation and Discussion Section.

This justifies the ideas of using randomly selected data and necessitates the experiments of the more phases. The next sections give their results respectively.

5.3. *Experimental Results of Phase 2*

5.3.1. *Fix-sized ensembles: 2+1 and 4+1*

For comparative reasons, the tussle for the opening positions in E built from members trained on data selected randomly from their respective subsets, is presented in the same sequence as section 5.1.

As can be seen, Figure 11 is remarkably different from its counterpart (figure 3). This trend continued throughout this phase. The most notable fact is that the

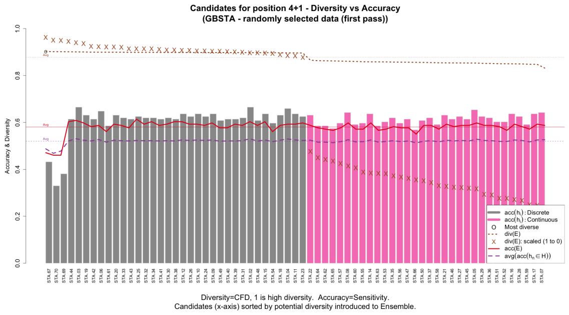

Figure 12: Phase 2: the diversity and accuracy results of 4+1 ensembles on randomly selected data GBSTA.

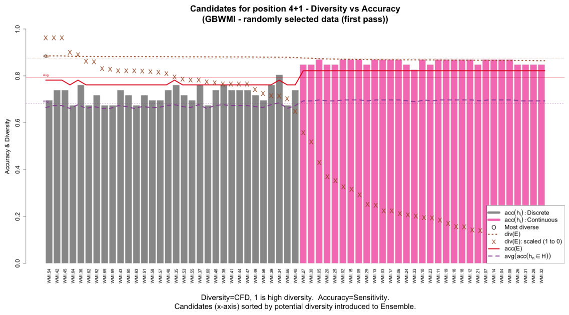

Figure 13: Phase 2: the diversity and accuracy results of 2+1 ensembles on randomly selected data GBWMI.)

accuracy $acc(h)$ is significantly higher than those models trained on the hand selected data. Diversity in ensembles $div(E)$ also remains high. With a few exceptions, where the $acc(h)$ falls below the default accuracy 0.5, so in principle, the candidates with $acc(h) >= 0.5$ would have made a positive contribution to ensemble E (compared to $avg(acc(h_n \in E))$).

When choosing the fifth member to build 4+1 ensembles, the results as shown by figure 12 indicate that the three most diverse candidates are the worst performers as individuals. Their accuracy is below 50% and they then drag down $acc(E)$ to a point where it is lower than $avg(acc(h_n \in E))$, hence these ensembles do not improve their collective accuracy at all. On the other hand, adding other candidates which are less diverse but more accurate into hybrid ensembles improve accuracy as $acc(E) > acc(h_n \in H = E)$. This means that the accuracy of ensembles if influence by both diversity and accuracy factors, not just diversity, nor accuracy itself.

The results of the hybrid ensembles built for GBWMI on randomly selected data, shown in figure 13, are even better than the above presented results for GBSTA data. It is clear that none of the individual members has poor accuracy and $acc(E)$ is always higher than $avg(acc(h_n \in E))$. The same pattern observed in figure 12 becomes even more obvious, which is, that $acc(E)$ for the most diverse but less accurate candidates(left half) is lower than $acc(E)$ for the least diverse but slightly more accurate candidates(right half).

Table 4: Phase 2: Statistics of the mean accuracies and variances of individual candidates induced from the randomly-selected GBWMI datasets with Discrete attributes, denoted by Z, and Continuous R.

	Z	R	Combined
Average	0.72	0.86	0.79
Standard Deviation	0.034	0.012	0.075

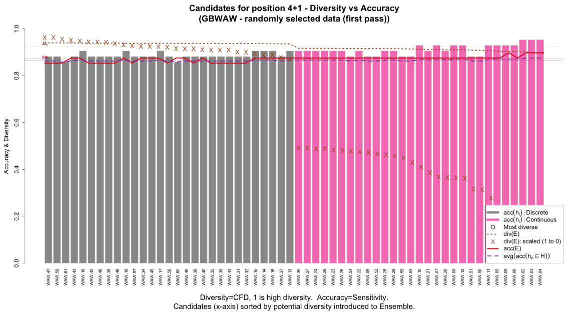

Figure 14: Phase 2: the diversity and accuracy results of 2+1 ensembles on randomly selected data GBWAW

An interesting observation is the clear divide between classifiers trained on different types of data. Table 4 shows that the numerical statistics of the mean accuracies and standard deviations of the two types of models induced from the datasets with discrete(Z) and continuous(R) attributes.

Close scrutiny of the results, shown in figure 14, of ensembles built with randomly selected data for GBWAW, reveals that while the overall $acc(E)$ remains absolutely high, the difference between $acc(E)$ and the mean accuracy of the members in $H = E$, $acc(h_n \in H)$ is tiny, it occasionally dips below $avg(acc(h_n \in H)$. This implies that there is a little gain achieved by the ensembles in terms of sensitivity accuracy, although the candidates are very accurate ($minacc(h) >= 82\%$) and quite diverse from the core members (the expert system and the most accurate candidate).

5.3.2. *Growing ensembles*

As in Phase 2 additional test data (True test) was designated and kept aside for testing the ensembles, apart from using the validation data to test the candidates. So the results presented in this subsection include the sensitivity accuracies: $acc(Validation)$ as well as

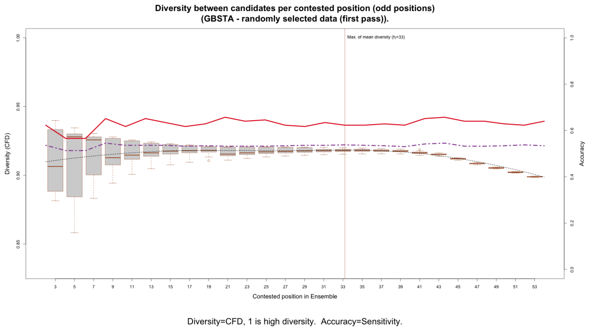

Figure 15: Phase 2: the diversity and accuracy results of growing ensembles on randomly selected validation and true test data for GBSTA. Note: there are two additional lines on this figure and all onwards ones: solid (blue) line with solid square markers is the mean sensitivity accuracy of hybrid ensembles on the true test data, and solid (orange) line with triangle markers is the mean sensitivity accuracy on the validation data.

$acc(TrueTest)$, obtained from both the validation data and True Test data respectively.

Again, in these experiment set-ups, the situations presented in the immediate above section are just some special cases when the size of ensemble is set to a fixed number.

It can be observed quite clearly from figure 15, for GBSTA, as E increases its size, the accuracies on validation and true test data: $acc(E(Validation))$ and $acc(E(TrueTest))$, are better than that of their individual members every time.

Further inspections found that the ensembles of 3 members are reasonably good on validation and true test data as their both accuracies are higher than the average of their members. However, when the ensemble size increases up to around 11 or 13, the accuracy of these ensembles drops, although it is still higher than the members' mean $acc(h_n)$. This is once again evidence that adding members to E based only on the diversity $div(E)$ they introduce to the ensembles will not necessarily result in a higher performing E.

Then, after this, the ensemble accuracy starts to increase again to even higher than $acc(E)$ when $N = 3$. Furthermore, after a certain point (about N=17) $acc(E(TrueTest))$ stabilises and adding more members does not change its accuracy. This means that there is no gain for building larger ensembles.

The results of the growing ensembles for GBWMI data, shown by figure 16, indicate that $div(E)$ reached a peak at position 3. Again E performed better than $avg(acc(h_n \in H))$ on both the Validation and True test data. Again the trend is that the results level-off after a while, which means that increasing the number of the members to E has not helped in increasing the accuracy of ensembles.

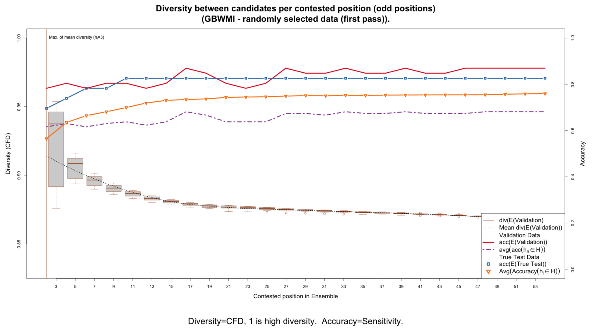

Figure 16: Phase 2: the diversity and accuracy results of growing ensembles on randomly selected validation and true test data for GBWMI.

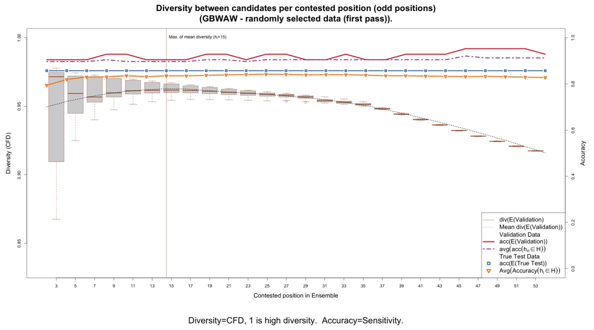

Figure 17: Phase 2: the diversity and accuracy results of growing ensembles on randomly selected validation and true test data for GBWAW.

The results of the growing ensembles for GBWAW, as shown on Figure 17, are similar to those shown on figure 16. At a point, the ensemble accuracy $acc(E)$ levels off, that is, there is no or little gain when adding new members into ensembles. Overall, all these ensembles perform very well consistently, with the average sensitivity over 85% on the test data, and specificity close to 100%.

5.3.3. Summary of the results of Phase 2

The results produced in all growing ensemble experiments for three counties are summarised in terms of the mean and standard deviation of the sensitivity measures and listed in Table 5.

Some interesting patterns are observed from the results obtained in Phase 2:

- In general, individual models trained with randomly selected data outperform those trained on manually selected data. Particularly on the first

Table 5: Statistical summary (mean,& standard deviation) for all ensemble results for Phase 2 on the true test data.

	$h_n \in E$		E	
Data	mean	stdev	mean	stdev
GBSTA	0.60	0.062	0.64	0.090
GBWAW	0.83	0.026	0.85	0.041
GBWMI	0.72	0.093	0.79	0.132

two counties, models are much better than their counterparts (shown in figures 3 & 4 and 11 & 13 ,) and usable for building ensembles.

- Consequently the performance of ensembles was also better and consistent over all the data subsets.
- The results indicate that the size of ensemble can have some influence on the ensemble accuracy $acc(E)$ at some ranges, but have little or no difference after a certain point. This suggests that it is not true that the bigger an ensemble is, the better it performs. There is a trade-off at some point.
- The results produced in Phase 2 clearly demonstrated that the randomly-selected data have a much better representation of the underlying problem than the hand-selected data.
- The accuracies on validation and true test datasets are mostly consistent and therefore validation performance is a good indicator for selecting ensembles that are more likely to be accurate and reliable on test data.

The experiments in Phase 2 were repeated in two more phases, i.e. Phase 3 and 4, with different data partitions to check if these results are consistent and reliable. Their results are found to be largely in line with the results obtained in Phase 2.

5.4. Summary of the results

Several points can be drawn from the results presented earlier.

- The results produced in Phase 2 (and two repeated phases) clearly demonstrated that the randomly-selected data have a much better representation of the underlying problem than the hand-selected data.
- The individual models are more sensitive to the change of data, on average, whilst the hybrid ensembles are robust due to the combination of hybrid models.
- On average, hybrid ensembles are more accurate and reliable than individual members if constructed appropriately, as $mean\{acc(E)\} > mean\{acc(h)\}$.
- Sometimes adding a more accurate $h_{candidate}$ did not improve $acc(E)$ - it actually reduced $acc(E)$ to a point where it fell below $avg(acc(h_n \in E))$.
- Sometimes adding a more diverse $h_{candidate}$ to E did not improve $acc(E)$ as much another $h_{candidate}$ based on a different criteria ($acc(h_{candidate}$) for example) would have.
- The results indicate that the size of ensemble can have some influence on the ensemble accuracy $acc(E)$ at some ranges, but have little or no difference after a certain point. This suggests that it is not true that the bigger an ensemble is, the better it performs. There is trade-off at some point.
- The accuracies on validation and true test datasets are mostly consistent and therefore validation performance is a good indicator for selecting ensembles that are more likely to be accurate and reliable on test data.

The results of all the ensembles are generalised in term of goodness and shown in Table 6. A good performance is determined by whether the average sensitivity measure $acc(E)$ of E (across its maximum permitted size) less one standard deviation is above 0.50 – the default accuracy, that is, $mean\{acc(E)\} - stdev(E) > 0.5$, poor, otherwise. However, a mean accuracy over the whole length of growing ensembles may vary up or down, i.e. oscillates a bit around the mean value, which is not surprising when the number of the members changes and the quality of the members varies.

Table 6: Ensemble accuracy in all phases for three counties.

Phase	GBSTA $acc(E)$	GBWMI $acc(E)$	GBWAW $acc(E)$
1	poor	poor	good oscillates
2+	good - with dip	good	good

6. Discussion

There are several interesting and important issues raised from this research, which include the data and choice of the data for training models, the expert system, machine induced models and ensembles. This section attempts to give some discussions on each of them in order to provide some guidelines .

6.1. Data Issues

The most prominent issue regarding the data is what is the best way to select data for training. This research used experienced human to select the data manually to begin with. The initial consideration was that the experienced human could apply their knowledge and experience to select the data that can represent the underlying problem better. But it was quickly discovered that was not the case, and then switched to random selection strategies.

Hand selected data vs Randomly selected data:

As the results produced in Phase 1 show that only one set of the three hand-selected data sets is reasonably representative as the generated decision trees and the ensembles performed reasonably well(with a mean sensitivity around 73%) but the other two failed to produce reasonable ensembles (with sensitivity only around 10% to 15%), although some individual classifiers performed well.

This is a sharp contrast to the randomly selected data sets, as shown by tables 5 and 6, the ensembles produced from the randomly selected data sets performed well or very well on the true test data, with the mean sensitivity between 64%, 85% and 79%, respectively on the true test data of three subsets.

These results are a clear evidence that randomly selected data better represented the problem domain than the manually selected data, although the manual selection was guided by the experienced human with an intention of preserving any annual patterns that might exist in the data (for example – during the autumn migration mistaken identification of very rare species is common – also some species are only found in the UK during particular periods of time) - so the relative poor performance of this data was a surprise.

In addition, considerable time and effort was spent in discretizing some continuous attributes, but the results demonstrated this effort was not paid off. One possible reason is that the number and width of the bins are not optimized. However, optimization is not an easy task as it involves many factors, such as distribution of input attributes and dependent output target, although there are some ways that could be tried. Nevertheless, our results indicate that in practice one should not try it in general unless there is a clear reason and a good method to do it.

6.2. The members of ensemble

The Expert System:

The Expert System could perform quite well (86% Sensitivity and 99% Specificity on the validation data for one county GBWAW), and consistently responsible for setting acc_{low}. However, the implementation was relatively simplified and could be extended/adapted with more sophisticated knowledge rules.

It is an interesting observation that conceptually the Decision Tree algorithm selected for the Machine Learning classifiers is very similar to the Expert System - although typically these algorithms produce many thousands of rules along seemingly random lines of enquiry compared to the rules in an Expert System.

Machine Learning Algorithms:

At the early stages of this work, other algorithms, e.g. kNN, Logistic Regression and Support Vector Machine(SVM) were tried, but were found not efficient or failed completely in dealing with the data of very high dimensionality and huge quantity. Then an equivalent implementation of C5.0 Decision Trees in R on a High Performance Computing Cluster (HPCC) was used, primarily for its efficiency and competitive accuracy. No comparison with other algorithms was attempted as there were no other implementations available on the HPCC.

It was noticed that Discrete and Continuous data sets were used to train models with the exact same procedure. It can be seen from figures 3 to 7 in Phase 1, the models are very different in most of cases. But, when the randomly selected data is used, there is no or little difference between them. However, close scrutiny coupled with domain knowledge reveals that the decision tree induced with the continuous attributes is too specific – common and rare birds are rejected out of hand without further evidence.

Hundreds of successful models were created and then several good ensembles were constructed.

6.3. The ensembles: accuracy, diversity and size

This research was set to develop a framework for building hybrid ensembles E and investigate what factors influence their performance. This study considered three factors: accuracy of members $acc(h)$, diversity $div(E)$ and the size of ensemble.

Accuracy and Diversity:

As mentioned before, these factors are coupled together, particularly the accuracy and diversity are closely tied, it is extremely difficult to isolate them and then examine their individual impact on accuracy of ensemble. In addition, there is no practical way to actively control/manipulate the change of accuracy and diversity, except for the size, which can be easily controlled. Therefore the investigations were practically conducted in a passive manner, i.e. using the models generated with whatsoever accuracy and diversity to build ensembles with the rules described in Section 3.1, including selecting candidates based on diversity and accuracy boundaries (acc_{low}, acc_{high}) on the selection of potential members.

The results of diversity measured among the classifiers show that these machine induced models are more similar to each other in such an extend that their diversity is not high enough to improve ensembles. This is not surprising as some studies such as Wang et al. (2001), already pointed out that the models generated with the same learning algorithm are statistically more dependent or similar than those by different learning methods. That is why this study introduce a methodologically different model – expert system, in a hope of boosting the diversity. This is justified by the facts that the highest diversity values are almost always achieved on the far left of the plots in all the figures because the expert system is more diverse from machine induced classifiers and hence taken as the first core member in any hybrid ensembles.

Constructive and Destructive Diversity:

Wang (2008) argued that diversity among the members should be measured in two ways: positive and negative, or namely constructive and destructive respectively, because diversity can be useful as well as harmful. But this phenomenon has not been observed in applications before and hence not paid enough attention in ensemble research. In this work, several examples were found, as shown in figures 3 and 4, where the accuracy of ensembles is worse than the average accuracy of the members, i.e. $acc(E) < acc(h_n \in E)$, which means that the members used in an ensemble actually work against each other to produce more wrong decisions than correct ones. In such a situation, the difference among must be destructive and hence this difference should be represented by a negative diversity measure.

Nevertheless, there is no definition found in literature to reflect negative diversity that causes the destructive effect on performance of an ensemble. Thus, there is a need to develop new measure for representing both positive and negative diversity, but it is beyond the scope of this study.

Size of Ensemble: It is quite clear that size does matter to a certain degree, but it then has no or little influence on the accuracy of ensemble after a certain point. So, in general, it is not the case that the larger an ensemble is, the better it performs. There is a trade-off and often this point is actually quite low, less than 19 in this study.

7. Conclusion

In this study, a hybrid expert ensemble framework has been proposed to combine an expert system and machine learning classifiers to identify reliable observations in BirdTrack data, collected through a citizen science project.

It has been implemented and tested on several data subsets selected in two ways, manually and randomly. Manual selection was intended to utilize the human's experience to select that data to better reflect the characteristics of the problem, such as, seasoning, species distribution, migrating patterns etc. for training. However, the experiments showed that these manually selected data sets are not representative for most cases, as the models trained on these data sets performed extremely poor on two of three test data subsets, acceptable on only one. Conversely, the randomly selected data sets have much better representations as the models trained on them generalised well enough to be considered as the candidates for building hybrid ensembles.

Due to the complexity and nature of species distribution, it was discovered that data for each county behaved differently and building a big ensemble for the whole country does not generalise well. Then a strategy of "divide-and-conquer" is adopted, that is, dividing the whole country into counties or regions and then building an ensemble for each of them. the experimental results demonstrated this strategy produced much accurate and reliable ensembles.

An expert system was built using the rules extracted from human experts' experience, with an intention to allow it work in the "same manner" as a human expert. It was tested on various data subsets and achieved reasonable performance. Although it varied considerably from county to county. It was then used as a core member when building hybrid ensembles. The experimental results indicated that combining an expert system with machine generated models brings greater diversity to an ensemble and boosted the performance.

In conclusion, this study has demonstrated that hybrid Expert ensembles are capable of identifying unreliable observations with mean sensitivity accuracy measured between 76% to 83% on three true testing data sets respectively, and specificity 99% to 100%. That means the hybrid ensembles are able to correctly identify the unreliable observations with an accuracy of around 80%, and 99% to 100% accurate for reliable observations.

Considering there are over 40 million observations in BirdTrack data to date, and rapidly increasing, it would be expensive and time-consuming to employ experienced humans to validate them all. With the technology developed in this work, it is possible to develop some hybrid ensemble systems to assist humans to do the work more efficiently and effectively, by at least filtering out the reliable records and then labelling the suspected unreliable ones for further inspections by human to begin with. There is certainly some room for further development and this kind of system may eventually work independently without human intervention.

Further work on the system can be done in several aspects, including

On data: collect and add more features such as the observer's profile, replace grid reference with numeric equivalent, and carry out cluster analysis on the data to find better grouping of the data, not just on geographical locations.

On Expert System: more experts' knowledge needs to be codified and represented with more rules, especially some local knowledge related to regions or counties. When more rules have been established, the expert system can certainly be improved and sub-expert systems added to address relatively poor generalisation on some counties.

On hybrid expert ensemble: several issues should be addressed, including: to explore new definition to measure both constructive and destructive diversity which would be more directly useful for building an effective ensemble; to devise better strategies for selecting suitable members to build more accurate ensembles, possibly based on a combination of accuracy and diversity, and to estimate the number of members needed for an ensemble.

It is worth noting that the framework developed in this research can be applied to other citizen science projects to validate the quality of big data or information contributed from the general public.

Acknowledgements

The authors would like to thank Dr. Andy Musgrove at the BTO in assisting with the data extraction and also the High Performance Computing service at the University of East Anglia (UEA) as the experiments in this study were partially carried out on their High Performance Computing Cluster.

References

Aldehim, G. and Wang, W. (2017). Determining appropriate approaches for using data in feature selection. *International Journal of Machine Learning and Cybernetics*, 8(3):915–928.

Anifowose, F., Labadin, J., and Abdulraheem, A. (2016). Hybrid intelligent systems in petroleum reservoir characterization and modeling: the journey so far and the challenges ahead. *Journal of Petroleum Exploration and Production Technology*, 7:251–263.

Bateman, I., Day, B., Agarwala, M., Bacon, P., Badura, T., Binner, A., De-Gol, A., Ditchburn, B., Dugdale, S., Emmett, B., Ferrini, S., Fezzi, C., Harwood, A., Hillier, J., Hiscock, K., Hulme, M., Jackson, B., Lovett, A., E., M., ., M. R., Sen, Siriwardena, G., Smith, P., Snowdon, P., Sunnenberg, G., Vetter, S., and Vinjili, S. (2014). UK National Ecosystem Assessment Follow-On Workpackage 3a: Economic value of ecosystem services.

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., and Shirk, J. (2009). Citizen science: A developing tool for expanding science knowledge and scientific literacy. *BioScience*, 59:977–984.

Bonter, D. N. and Cooper, C. B. (2012). Data validation in citizen science: a case study from project feederwatch. *Frontiers in Ecology and the Environment*, 10(6):305–307.

Bowser, A. and Cooper, C. (2017). The state of the data in citizen science. *Biodiversity Information Science and Standards*, 1:e20370.

Cervantes, B., Monroy, R., Medina-Prez, M. A., Gonzalez-Mendoza, M., and Ramirez-Marquez, J. (2018). Some features speak loud, but together they all speak louder: A study on the correlation between classification error and feature usage in decision-tree classification ensembles. *Engineering Applications of Artificial Intelligence*, 67:270 – 282.

Ciarratano, J. and Riley, G. (2005). *Expert Systems Principles and Programming*. Thomson Course Technology.

Couzens, D. and Nurney, D. (2013). *Birds ID Insights*. Bloomsbury.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems*, pages 1–15. Springer.

Kowalski, J., Krawczyk, B., and WoÅniak, M. (2017). Fault diagnosis of marine 4-stroke diesel engines using a one-vs-one extreme learning ensemble. *Engineering Applications of Artificial Intelligence*, 57:134 – 141.

Kuhn, M., Weston, S., Coulter, N., Culp, M., and Quinlan, R. (2015). C5.0 decision trees and rule-based models.

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.

Lewandowski, E. and Specht, H. (2015). Influence of volunteer and project characteristics on data quality of biological surveys. *Conservation Biology*, 29(3):713–723.

Liang, D., Tsai, C.-F., Dai, A.-J., and Eberle, W. (2018). A novel classifier ensemble approach for financial distress prediction. *Knowledge and Information Systems*, 54(2):437–462.

Lukyanenko, R., Parsons, J., and Wiersma, Y. F. (2016). Emerging problems of data quality in citizen science(editorial). *Conservation Biology*, 30(3):447–449.

Partridge, D. and Krzanowski, W. (1997). Software diversity: practical statistics for its measurement and exploitation. *Information and software technology*, 39(10):707–717.

Perikos, I. and Hatzilygeroudis, I. (2016). Recognizing emotions in text using ensemble of classifiers. *Engineering Applications of Artificial Intelligence*, 51:191 – 201. Mining the Humanities: Technologies and Applications.

Rayana, S. and Akoglu, L. (2016). Less is more: Building selective anomaly ensembles. *ACM Trans. Knowl. Discov. Data*, 10(4):42:1–42:33.

Richards, G. and Wang, W. (2012). What influences the accuracy of decision tree ensembles? *Journal of Intelligent Information Systems*, 39(3):627–650.

Vinicombe, K., Harris, A., and Tucker, L. (2014). *The Helm Guide to Bird Identification*. Bloomsbury.

Wang, W. (2008). Some fundamental issues in ensemble methods. In *IEEE International Joint Conference on Neural Networks, 2008. (IEEE World Congress on Computational Intelligence)*, pages 2243–2250.

Wang, W., Partridge, D., and Etherington, J. (2001). Hybrid ensembles and coincident-failure diversity. In *Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on*, volume 4, pages 2376–2381. IEEE.

Wiggins, A. and He, Y. (2016). Community-based data validation practices in citizen science. In *In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW'16)*, pages 1548–1559, San Francisco, USA.

Wiggins, A., Newman, G., Stevenson, R. D., and Crowston, K. (2011). Mechanisms for data quality and validation in citizen science. In *"Computing for Citizen Science" workshop at the IEEE eScience Conference*, Stockholm, Sweden.

Yousefnezhad, M., Reihanian, A., Zhang, D., and Minaei-Bidgoli, B. (2016). A new selection strategy for selective cluster ensemble based on diversity and independency. *Engineering Applications of Artificial Intelligence*, 56:260 – 272.

Appendices

Appendix A. Important Attributes in BirdTrack Data

Table A.7: Important Features or Attributes. Z=Discrete, R=Continuous, where both are available a suitable conversion function was provided.

Type	Feature Name	Behaviour	Description
Logical	`all_obs_reported`	Z	Observer reported all birds or just highlights
Character	`county_code`	Z	County Code
Real	`distance_to_coast`	R	Approximate distance to coast
Real	`duration_hrs`	R	Hours spent watching birds for this list
Integer	`easting`	R	British National Grid Easting. See `gridref`.
Character	`gridref`	Z	Ordinance Survey Grid Reference for centroid of Location. Site may extend beyond square kilometer but the assumption is that all observations occurred within square kilometer.
Logical	`has_activity`	Z	Observer commented on activity bird was engaged in
Logical	`has_age_sex`	Z	Observer indicated age or sex of bird
Logical	`has_comments`	Z	Observer provided additional comments. Potential bias - it is not clear whether the user entered the comment before or after the observation is identified as "unusual".
Logical	`has_direction_of_flight`	Z	Observer noted direction of flight
Logical	`has_habitat_notes`	Z	Observer made notes about habitat
Logical	`has_proof_of_breeding`	Z	Observer noted the bird was (potentially) breeding
Logical	`has_sensitivity`	Z	Observer requested data not to be publicly available
Logical	`has_significant`	Z	Observer notes observation is remarkable (for any reason, including personal reasons)
Integer	`how_many`	R,Z	How many individuals of indicated species seen
Integer	`list.week`	Z	Week of the observation
Integer	`list.year`	Z	Year of the observation
Logical	`mobileapp`	Z	yes=list added via Mobile Application, no=added via Website
Integer	`northing`	R	British National Grid Northing. See `gridref`.
Integer	`num_lists`	R, Z	Number of lists submitted by observer, to date
Integer	`num_partial_lists`	R, Z	Number of lists (with partial observations) submitted by the observer, to date
Integer	`pvkey`	Z	Unique identifier for an observer
Integer	`species_code`	Z	Species specific identifier
Character	`verification_status`	Target	0=Invalid, 1=OK, other=In Query