Handling Uncertainty in Citizen Science Data: Towards an Improved Amateur-based Large-scale Classification

Manuel Jiménez[a,*], Isaac Triguero[a], Robert John[a]

[a] *The Automated Scheduling Optimisation and Planning Research Group, School of Computer Science, University of Nottingham, Nottingham NG8 1BB, United Kingdom*

Abstract

Citizen Science, traditionally known as the engagement of amateur participants in research, is showing great potential for large-scale processing of data. In areas such as astronomy, biology, or geo-sciences, where emerging technologies generate huge volumes of data, Citizen Science projects enable image classification at a rate not possible to accomplish by experts alone. However, this approach entails the spread of biases and uncertainty in the results, since participants involved are typically non-experts in the problem and hold variable skills. Consequently, the research community tends not to trust Citizen Science outcomes, claiming a generalised lack of accuracy and validation.

We introduce a novel multi-stage approach to handle uncertainty within data labelled by amateurs in Citizen Science projects. Firstly, our method proposes a set of transformations that leverage the uncertainty in amateur classifications. Then, a hybridisation strategy provides the best aggregation of the transformed data for improving the quality and confidence in the results. As a case study, we consider the Galaxy Zoo, a project pursuing the labelling of galaxy images. A limited set of expert classifications allow us to validate the experiments, confirming that our approach is able to greatly boost accuracy and classify more

*Corresponding author at: The Automated Scheduling Optimisation and Planning Research Group, School of Computer Science, University of Nottingham, Nottingham NG8 1BB, United Kingdom.

Email address: manuel.jimenezmorales@nottingham.ac.uk (Manuel Jiménez)

images with respect to the state-of-art.

Keywords:
Citizen Science, Classification, Astroinformatics, Galaxy Morphologies, Uncertainty, Data analysis

1. Introduction

Connectivity is promoting the emergence of a great potential amongst virtual communities of people that share a common goal. In some cases, this goal may consist of making a significant contribution towards the solution of a complex scientific problem. Whereas, in the past, the analysis of these problems used to be restricted to a group of experts in the subject, today this is difficult when the processing of large amounts of data is required. In this context, Citizen Science refers to the development of scientific research assisted by amateur volunteers from the general public [15]. As a form of crowdsourcing [13], this practice is re-emerging, engaging the crowd in helping researchers complete high time-consuming tasks for which no reliable automatic procedures are available yet, for example, labelling of images [45], detection of patterns in graphic data [49], or transcription of handwritten texts [22].

Here we deal with classification problems in Citizen Science, which generally aim at the classification of huge collections of images according to a number of classes. These classes capture the interest of a specific research field, and identifying them correctly is the target of the participants. This sort of project involves, for instance, the recognition of structures in cell images [35], animal species in images taken in the savannah [2], or types of storms in actual data taken from meteorological satellites [23]. Amongst others, the nascent discipline of astroinformatics [3] has greatly benefited from the analysis of astronomical data in multiple projects, providing data analysis at a scale never reached in the past [31, 37, 6]. Nevertheless, many challenges are raised when the maximal profit of this large-scale analysis is desirable, regarding aspects such as the best use of expert classifications [44] or participants' engagement in this type of

voluntary scientific contribution [42].

Citizen Science has also attracted the attention of data scientists. Research focused on the mining of data using an *off-line* approach, that is, the study of results once the project has finished. They have tested the capabilities of Data Mining (DM) and Machine Learning (ML) techniques, aiming the replication of amateurs' performance [5, 18] or the training of ML algorithms for a certain problem [38, 7]. Moreover, ML implementations are also being used for optimising amateurs' endeavours through the course of the project, following an *on-line* approach. This other framework encompasses the progressive training of new participants as they acquire experience in the problem, or the interaction between a ML classifier and new labelled data as it is generated by project participants [49].

Despite this, Citizen Science still arouses scepticism within the research community [39]. Even though it offers possibilities for research not possible to accomplish by experts alone, it is not universally accepted as a valid research method [10]. The reasons lie in the quality of results, which are often questioned because of several drawbacks involving the prevalence of biases and lack of accuracy and validation [28]. Amateurs participating in Citizen Science projects exhibit a wide range of skills, and it is not guaranteed they hold any background in sciences or research. Moreover, there is always some degree of uncertainty in classification problems, which usually tend to bring additional vagueness in the definition of the classes (type of birds, shape of galaxies, patterns in a graph, etc.). Consequently, depending on the problem and participants' expertise with the classification task, the confidence through amateur-labelled data varies and Citizen Science results thus hold an intrinsic uncertainty.

The study of classification problems with uncertain labels has been developed adopting several approaches [12, 30]. Nonetheless, when the uncertainty comes from a set of independent judgements on the object being classified, fuzzy logic provides a very suitable framework for a thorough study of this uncertainty [27]. Areas such as multi-criteria decision making and multi-expert decision making address the problem of providing a final decision when a set of indepen-

dent judgements are available [21, 46]. Several aggregation methods have been widely studied through the specialised literature, aiming to use a set of experts' individual preferences in such a way that reflects the properties contained in all individual contributions [14, 47]. However, this kind of approach has not been extended when there is available a great number of non-experts opinions with a widespread uncertainty in their final decisions. Moreover, classification problems covered by Citizen Science projects tend to produce additional vagueness related to the definition of the classes to be identified, a disparity either in the total number of votes received or in the confidence of amateur classifications, etc.

To the best of our knowledge, the enhancement of Citizen Science results by using this kind of methods has not been fully addressed yet. In our preliminary study [25], we started investigating the potential and the issues derived from the employment of these results with two simple data transformations. In this work, we propose a novel approach that, based on our previous findings, uses the data produced in Citizen Science projects that deal with classification problems. We present a method to aggregate information about the prevalent uncertainty in this sort of data. We first identify three sources of uncertainty in Citizen Science data that we address separately by a set of transformations that aim to enrich the original data. Then, we employ a hybridisation strategy that explores the most suitable combination of these individual transformations, providing more confident and accurate classifications. We eventually pursue a refinement of results, so that, they became more trustworthy and maximise the utility and outreach of Citizen Science projects.

We consider as a case study the first edition of the Galaxy Zoo (GZ1) project [32], one of the very first successful implementations of Citizen Science using the Internet. GZ1 finished classifying nearly one million galaxy images with the help of more than 200,000 volunteers. However, these results did not consider at that time a substantial part of the information stored in the original data about the uncertainty in amateur classifications. Making an integrated use of the same original GZ1 dataset, our approach is able to provide more accurate

classifications for a greater number of galaxies, improving the state-of-art of the problem.

This document is organised as follows. In Section 2, we extend the background on Citizen Science and the management of the uncertainty with fuzzy logic. In Section 3, we introduce our approach for the improvement of Citizen Science data. Section 4 presents the set of experiments that test our method along with a discussion of results. Finally, in Section 5 we draw some conclusions and outline possible directions for future work.

2. Background

In this section, we further introduce the main materials covered in this work for a better comprehension. In Section 2.1, we first explain in more detail different aspects around the running of Citizen Science projects and review current trends in the specialised literature. After this, in Section 2.2 we take a deeper look at the field of fuzzy logic as a promising resource for the improvement of Citizen Science results.

2.1. Citizen Science: A brief overview

Citizen Science has been a common practice for many years. This form of *citizen* support to science developed by volunteers goes back in time to the eighteenth century. In those days, a few amateurs started making small but significant contributions by reporting observations about meteorology and ornithology [34]. Nowadays, the great advances in the Internet and Information Technologies have broadened the ways volunteers can develop these research-related activities, to the point that Citizen Science is being re-discovered by the scientific community as a valuable resource [40]. An increasing number of projects engage day by day significant numbers of individuals through the Internet in collecting and/or analysing data, with the support of many institutions from research and academia. The Zooniverse[1] initiative is one of the main plat-

[1] `http://www.zooniverse.org`

forms for Citizen Science project development and management [41]. Currently, Zooniverse hosts more than 80 projects in topics such as space sciences, ecology, medicine and humanities, directing the joint effort of more than a million participants [20]. This has led to the publication of more than 250 scholarly articles[2], validating the utility of Citizen Science for today's research.

There is a solid body of works devoted to the study of Citizen Science as a social phenomenon, emphasising different aspects such as motivation of volunteers, challenges towards acquiring real research status, or its future prospects [15, 9, 40, 17, 10]. A shared claim within these works is the latent potential in the crowd as a valuable resource that should not be neglected by scientific community. Nonetheless, two main concerns are raised by scientists: a generalised lack of accuracy and a proliferation of biases within the data coming from Citizen Science projects [28]. To overcome this, it has become crucial the development of proper tools for improving data accuracy, control and minimise the impact of biases, and validate final results. These issues have been addressed from several approaches [1, 48, 11, 8]. However, this body of works focuses on Citizen Science projects in the context of ecological sciences, which aim for data collection from natural environments at a massive scale. They ignore the difficulties covered in this work that arise when the target is the processing of data by large amounts of people. This is the case, for example, in projects coping with classification of images.

Citizen Science projects usually involve one particular task around the processing of some sort of *raw* data. Once the project is released, participants interested in taking part are invited to complete the task, developing genuine data analysis. For a great number of projects, this has consisted of the classification of large collections of images. After some training is provided, amateurs are asked to classify the images displayed in the project website by choosing amongst a set of categories. These categories often hold a set of main classes, which get the major part of the votes and comprise the target of the classifica-

[2] A complete list of references can be found at `http://www.zooniverse.org/publications`.

tion problem. In addition, it is commonly offered a *Don't Know* (DK) category, useful in case no class is clearly distinguishable and that ensures any image gets a vote every time it is shown. When the project is closed down, all clicks conveniently recorded in a database are made available to a team of experts in the problem for their follow-up study. This data normally includes the count of votes for each of the classes offered to participants, and not a final label for each of the objects in the original dataset, as it is shown in Table 1. Therefore, a suitable analysis of this data is key at this point to extract good results from the project. However, a thorough study of this problem from the data science perspective remains unexplored.

Image ID	***Votes***	***Class 1***	***Class 2***	$\cdots$	***Don't Know***
0152948451	58	0.310	0.414	$\cdots$	0.052
0152863349	14	0.643	0.214	$\cdots$	0.071
0152878152	33	0.000	1.000	$\cdots$	0.000
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
0152721030	19	0.316	0.263	$\cdots$	0.263

Table 1: Typical look of a dataset recorded in the course of a Citizen Science project that involves image classification. Each row contains the information for each of the examples: the image ID, total number of votes received by the project participants, and the proportion of votes corresponding to the set of classes in the problem, including the *Don't Know* (DK) category.

Data science research applied to Citizen Science has mainly been dedicated to the exploitation of the data after the projects are concluded. This kind of off-line approaches enable the mining of the resulting data along with additional information related to amateurs' performance, available expert knowledge on the problem, and other statistics about the running of the project. To date, several works have been focused on emulating amateur classification skills by using ML algorithms, aiming to facilitate an automated analysis of images in diverse contexts. This goal has been achieved by taking as input either a set

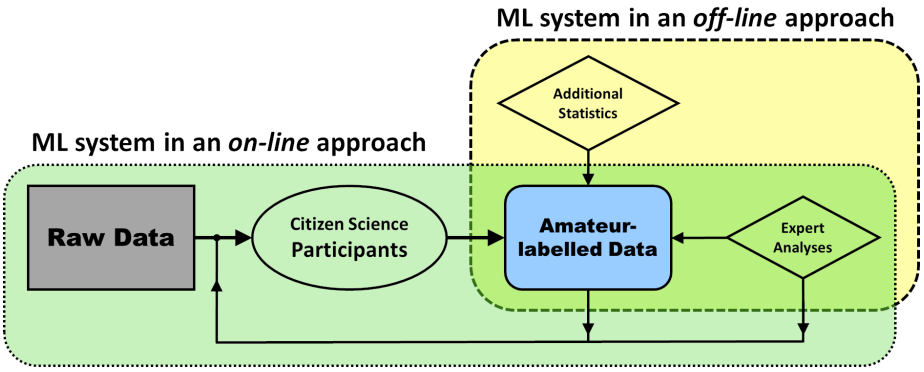


Figure 1: Two potential uses of ML within the Citizen Science workflow. *On-line* approaches take advantage of the synergy between experts and the training of new amateurs, learning from the amateur-labelled data. In contrast, *off-line* approaches aim to learn from the aggregation of amateur-labelled data, expert analyses and additional information available once the project has been concluded.

of features extracted from the image [5, 29], or the whole image within a deep learning approach (that performs its own feature extraction) [18]. It has been shown that ML classifiers can achieve similar results to those obtained by a group of amateurs, when these algorithms are trained using Citizen Science data [7]. However, these approaches do not address the intrinsic uncertainty, and tend also to replicate the biases present in the data. In addition to off-line approaches, on-line settings have recently been developed for optimising the interaction between humans and machines through the course of the projects [26, 16]. These approaches involve ML systems that deal with the training of participants as they acquire expertise in the problem, the management of expert classifications, and the synergy between amateurs, experts and automated classifiers [49]. The operation of both kind of approaches is outlined in Figure 1, where we highlight the interrelation amongst both potential roles of ML in reinforcing Citizen Science outcomes.

In the present work, we opt for an off-line approach that targets the inherent uncertainty in projects that tackle classification problems. Our aim is to help

experts increase their accuracy and confidence in amateur-labelled data in order to improve the outcomes of this kind of projects. To do so, our approach ensures an aggregation of information concerning this uncertainty that, as a form of data pre-processing, ensures a better use of this data for either research or the training of ML algorithms.

2.2. Fuzzy logic for handling uncertainty: A promising resource for improving Citizen Science data utility

Fuzzy logic tackles, amongst multiple other subjects, the various forms of uncertainty exhibited in a varied range of problems. As *ambiguity* and *vagueness* (or *fuzziness*) we identify, respectively, the lack of specificity when a set of choices is available, and the difficulty of making sharp or precise judgements in real-world problems [27]. These two concepts are deeply intertwined in Citizen Science, since usual activities required to amateurs frequently involve unspecified tasks such as the crisp classification of vague classes, transcription of ambiguous information, or identification of patterns. On the one hand, this kind of tasks bring themselves some level of uncertainty within their definition; on the other hand, we eventually count with a set of independent opinions. These opinions have to be aggregated in the most proper way to get the best results and then maximise the utility of Citizen Science projects.

Multi-criteria and Multi-expert decision making are well-studied categories of problems concerned with finding the best choice when a set of alternatives is available [43, 19]. Eventually, an aggregation method is needed to combine individual criteria into a final decision, which is expected to contemplate all individual contributions. In data coming from Citizen Science projects we often encounter this scenario, where there is available a set of opinions. However, while fuzzy models for decision making normally use information from a reduced number of experts on the problem, in the case of Citizen Science data the uncertainty is more extreme for two reasons: firstly, amateurs (in contradistinction to experts) hold a wide range of backgrounds and varying expertise on the task, meaning more vagueness in their opinions; second, the number of

judgements that need aggregation is much larger than in other typical group decision making problem. For instance, standard medical decision making has modelled the aggregation of $\sim$50 experts [21], whereas a typical Citizen Science project engages up to hundreds of thousands of participants, each one providing several tens of opinions about a set of objects.

These approaches represent a valuable initial framework for studying better use of Citizen Science data. A wide range of uncertainties is pervasive either through the problem definition, amateurs' set of judgements, and in the process of aggregating these judgements to reach a final classification. Depending on the nature of the problem addressed, results provided by amateur participants can be aggregated using expert knowledge in the subject to take advantage of all resources available. Pursuing this target, here we propose a way for aggregating additional information about the uncertainty in the voting process that, despite its simplicity, is able to improve current results.

3. A method for handling uncertainty in Citizen Science classification

In this section, we present our approach for handling the uncertainty spread within Citizen Science data. We consider the whole dataset obtained after the project has finished collecting votes from participants, taking an off-line approach. Firstly, in Section 3.1 we introduce basic notation and motivate the adequacy of the method by distinguish three types of uncertainty present in this sort of data. Then, in Section 3.2 we present a set of mathematical transformations that aims to leverage each of these uncertainty types. After this, in Section 3.3 we explain a hybridisation strategy that explores the best way to concatenate the three transformation stages in order to get the most convenient aggregation procedure.

3.1. Motivation: Three sources of uncertainty within Citizen Science data

In this section, we introduce the basic problem related to the employment of Citizen science results as well as some notation about Citizen Science data taken

in an off-line approach. Then, we provide brief explanations of the different ways the uncertainty is encountered within the data. This makes easier the later comprehension of the method.

This work focuses on Citizen Science projects that signify a valuable aid for some scientific research in solving a certain classification problem. The classification task, which is the core goal of the project, tends to involve the identification of a few classes across huge collections of images. However, as explained above, the task is developed with the help of a myriad of amateur participants. Hence, the output is not a final label for each of the images released during the project running but a variable set of independent amateur votes. Using the data generated by this process, here we propose a better use of these results adopting an off-line approach and exploring how to leverage information about the uncertainty in amateur votes that is able to improve the quality of final classifications.

In order to facilitate the subsequent data analysis, amateur votes are usually converted into scores, which are numbers in the unit interval calculated dividing the number of votes in each category by the total number of votes received by the example. Thus, let $\mathbf{V} = (v_1, v_2, ..., v_C)$ be the vote vector for an instance in the dataset, containing the votes for each of the categories defined in the problem, with C the number of categories and $N = \sum_{i=1}^{C} v_i$ the total number of votes received by that object. We get the score vector $\mathbf{X} = (x_1, x_2, ..., x_C)$ by computing $x_i = \frac{v_i}{N}$, for $i \in \{1, 2, ..., C\}$. The score vector is typically used to obtain a final classification for the object by simply applying a threshold: the category which score is greater or equal than the threshold is assigned to that example. This procedure allows the expert to adjust the confidence in the classification: the higher is the threshold applied, the larger is the consensus amongst amateurs who labelled that object, and objects holding a greater consensus are expected to be assigned more accurate classifications. However, the selection of the threshold is arbitrary, and even more importantly, it does not take into account the total number of votes, N. On the one hand, all objects which scores do not reach the threshold are left unlabelled (*uncertain*), mak-

ing the process ineffective as we require higher confidence in the classifications. On the other hand, examples with similar scores may hold a totally divergent number of votes N. So we are neglecting a hidden disparity in confidence.

The main issue derived from the employment of Citizen Science data is the prevalent uncertainty when a group of people provides a set of judgements about the same object. Amateur participants are not expected to agree in their classifications, and final labels depend on how this disagreement is handled. Additionally, we often encounter variability in the total number of votes received by the example, N. Our target here is to refine this amateur-labelled data in order to obtain better classifications and improve both the number of objects classified by applying a threshold as well as the quality of these classifications. We distinguish three different sources of uncertainty within the data:

- We refer to Inherent Uncertainty (IU) as the uncertainty due to the variation across amateurs' votes. Given an example displayed in the website, each participant is asked to classify it by clicking in the most appropriate category according to their opinion at the time. Therefore, the final outcome is not a classification but a record of votes for each of the categories, which spread tells us about the IU in that object. In the case all participants have voted for the same category, this class holds a 1.0 score and then the example presents zero IU. Conversely, if the votes are equally split across the categories, with scores equal to $\frac{1.0}{C}$, the IU reaches its highest value accounting for the greatest uncertainty.

- We denote as Measured Uncertainty (MU) the uncertainty directly quantified by the DK category. This option is normally offered as a form of ensuring every example gets a vote every time it is shown to a participant. This count of votes represents a measure of the uncertainty in the classification: as one object holds a greater number of DK votes, v_{DK}, it is expected to entail more ambiguity in its labelling. Hence, an example with $v_{DK} = 0$ ideally holds zero MU, getting bigger as v_{DK} takes on larger values.

- Lastly, we refer to Level of Confidence (LC) as the uncertainty caused by the variability in the total number of votes, N, received by each of the examples in the dataset. This quantity often follows an uneven distribution, being able to provide an estimation of the confidence in the classifications with respect to the whole set of examples: given an example, the higher is N in comparison with the rest of the set of objects (taking a metric, for instance, the mean number of votes, μ_N), the greater is our confidence in the set of scores for that example. Consequently, for scores similarly spread through the categories of the problem, the LC informs about the more or less confidence we can expect in regards to each particular example.

The three sources of uncertainty are inevitably intertwined. The MU is part of the IU, which accounts for the spread of the votes through the whole set of categories, including the DK votes. The LC, in turn, is codified in the IU as well, since we can trust a finer variability in the scores given an example as more votes are available, that is, as N reaches greater values. Here we do not aim to study these concepts in depth. We only set a concise conceptual framework for the explanation of the method.

3.2. Three transformations for data refinement

In the following subsections, we explain the basis of the proposed method, consisting of three independent mathematical transformations to be applied on the original scores. These transformations are intended to aggregate information about the uncertainties summarised above and not present per se in the set of scores obtained from amateur votes. For the sake of clarity, we label each one with a number tag (not related to any order or importance) and explain their application over the example data presented in Table 1. The method takes as input the whole set of vote and score vectors, $\mathbf{V}$ and $\mathbf{X}$, respectively, for each example in the dataset, and provides a modified score vector. Using the new scores, we can apply a threshold to assign a final class to the example, as it was explained above.

3.2.1. Normalisation: Reinforcement of main classes

The first transformation ($\{1\}$) consists of the normalisation of a subset of the scores. In Citizen Science projects dealing with classification problems, we commonly find that some classes within the options available for voting covers the major part of the examples. These so-called *main classes* hold a greater importance with respect to the rest and represent the target of the problem, that is, to classify the sample according to these few main classes. For example, participants may be asked to recognise either shapes of celestial objects, patterns in a graph, or types of animals in a picture of the savanna, all of these previously defined as canonical types. In addition, other secondary classes are offered, corresponding to minority (less common or rare) classes in the problem or a *Don't Know* response for the extreme cases in which the amateur is not able to decide. These secondary classes may be of interest for other problems. In this work we are focusing on the improvement in the classification of the main classes. Once the scores are computed, the minority classes tend to obtain negligible scores and therefore do not reach the threshold for the vast majority of the examples. However, these secondary scores contribute to lower the main classes scores, complicating the classification with a threshold. Hence, the normalisation of the main scores is intended to remove the "noise" due to votes received by secondary classes. We also obtain a representation of the IU restricted to the target classes of the problem and independent of the total number of votes received by the example, N: all instances with equal proportion of votes in the main categories are assigned identical scores after the normalisation.

Let $\mathbf{X} = (x_1, x_2, ..., x_C)$ be the whole score vector, we select the main scores, getting a reduced score vector $\widehat{\mathbf{X}} = (x_1, x_2, ..., x_M)$, with M the number of main categories ($M < C$). Once we have $\widehat{\mathbf{X}}$, the normalised score vector $\widehat{\mathbf{Z}} = (z_1, z_2, ..., z_M)$ is obtained performing a usual normalisation as shown in Equation 1, for $i \in \{1, 2, ...M\}$.

ID	***N***	***C1***	***C2***	***Norm. C1***	***Norm. C2***
0152948451	58	0.414	0.310	0.572	0.428
0152863349	14	0.643	0.214	0.750	0.250
0152878152	33	0.000	1.000	0.000	1.000
⋮	⋮	⋮	⋮	⋮	⋮
0152721030	19	0.316	0.263	0.546	0.454

Table 2: Normalised scores for the two main classes ***C1*** and ***C2*** in the example data presented in Table 1. Each row includes the image identification (***ID***), total number of votes (***N***), original scores and normalised (***Norm.***) scores.

$$z_i = \frac{x_i}{\sum x_i} \tag{1}$$

The normalisation of the main scores ensures that $\sum_{i=1}^{M} z_i = 1$ for every example. This develops as well a *cleaning* of the main scores for a later aggregation of information about the MU and LC by the two other transformations. Taking as example the data presented in Table 1, the normalised scores for this data are shown in Table 2, assuming this is a problem with two main classes: *C1* and *C2*.

3.2.2. DK votes shift: evaluation of Measured Uncertainty

The second transformation ($\{2\}$) modifies the main scores using the information held in DK votes. It aims to leverage the MU of the example by introducing a shift that favours one particular class and penalises the rest. In projects dealing with classification, we usually find an asymmetry in the main classes: one class is harder to identify than the rest. This occurs, for example, when the overall quality of the images is deficient because multiple factors (images of natural environments affected by weather conditions, space images that depend on the distance, etc.), or biases emerge in amateurs' skills (for instance, due to a unequal number of examples displayed for each of the classes, so more repeated classes become easily recognisable). In this case, the *Don't Know* category quan-

tifies to some extent this uncertainty, when the number of DK votes keeps low with respect to the total amount of votes allocated to the main classes. The shift targets this "imbalance": adds confidence to examples with moderate MU. However, it vanishes for instances with high number of DK votes and prone to hold high IU and therefore be uncertain.

For the calculation, two quantities are incorporated. One is related to the example at hand and another one is taken as a global measure: the number of DK votes for the example, v_{DK}, along with the average number of DK votes across the entire dataset, μ_{DK}. These are combined for the computation of a quantity ϵ, as it is shown in Equation 2. Once the shift ϵ is calculated for each of the instances, it is added to the score of the selected class. The remaining main scores are equally subtracted the proportional part of ϵ, depending on how many main categories there are. Being x^* the favoured class in $\widehat{\mathbf{X}} = (x_1, x_2, ..., x^*, ..., x_M)$, the shifted score vector $\widehat{\mathbf{W}} = (w_1, w_2, ..., w_M)$ is computed as it is shown in Equation 3, for $i \in \{1, 2, ...M\}$.

$$\epsilon = \frac{\alpha \cdot \mu_{DK}}{\beta + v_{DK}} \tag{2}$$

$$\begin{cases} w_i = x_i + \epsilon_i, & \text{for} \quad x_i = x^* \\ w_i = x_i - \frac{\epsilon_i}{M-1}, & \text{in other case} \end{cases} \tag{3}$$

Additionally, this transformation uses two parameters that modulate the modification introduced to the scores, which can be adjusted depending on the nature of the problem. The parameter α works as a factor that regulates the influence of the shift over the original scores. The parameter β is added to the count of DK votes, v_{DK}, for the calculation of the shift. These two parameters are optimised by testing a set of pair of values and assessing the modified scores with expert classifications as ground truth. Ultimately, we also restrict the range of application of the transformation: we discard examples with maximum IU (all scores equal to $1.0/M$) and zero IU (one category holds 1.0 score and the rest 0.0 score). The optimisation of parameters and range of application is illustrated within the experiments in Section 4.

ID	*N*	*DK*	v_{DK}	ϵ	*C1*	*C2*	*S. C1*	*S. C2*
0152948451	58	0.052	3	0.038	0.310	0.414	0.348	0.376
0152863349	14	0.071	1	0.075	0.643	0.214	0.718	0.139
0152878152	33	0.000	0	–	0.000	1.000	0.000	1.000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0152721030	19	0.263	5	0.025	0.316	0.263	0.341	0.238

Table 3: Shifted scores for the two main classes ***C1*** and ***C2*** in the example data presented in Table 1. Each row includes the image identification (***ID***), total number of votes (***N***), DK score (***DK***), DK number of votes ($\boldsymbol{v_{DK}}$), the value of the shift for the instance ($\boldsymbol{\epsilon}$), and original and modified (***S.***) scores. The DK votes are computed by multiplying the DK scores by the total number of votes. The values of $\boldsymbol{\epsilon}$ are obtained from Equation 2 with $\alpha = 0.1$, $\beta = 1$ vote, and $\mu_{DK} = 1.5$ vote. The original scores for ***C1*** and ***C2*** are obtained from Equation 3 with $M = 2$, and being ***C1*** assumed as the favoured class of the two-main-classes problem.

Considering again the example data in Table 1, we demonstrate the application of this transformation $\{2\}$ in Table 3. As an example, we use the values $\alpha = 0.1$, $\beta = 1$, and $\mu_{DK} = 1.5$ vote, as adjusted for this particular problem to compute the values of ϵ for each of the instances. Again, we assume there are two main classes, and *C1* is the favoured class. Consequently, the shift values ϵ are added to *C1* scores and subtracted from the *C2* scores.

3.2.3. Votes boost: Addition of confidence to highly-voted examples

The third transformation ($\{3\}$) modifies the main scores employing the information present in the LC. In this case, each score is incremented using the distribution of the number of votes for the class across the entire dataset. Again, we are only interested in the main classes of the problem. However, unlike transformations $\{1\}$ and $\{2\}$, this boost always increases the scores in accordance with the total number of votes received by the class.

The scores are modified as follows. In the first place, the number of votes for each of the main classes, v_i, are expressed in standard units as it is shown

in Equation 4, taking the mean, μ_i, and standard deviation, σ_i, over the entire dataset for the selected class. After this, the score converted in standard units is weighted by the sigmoid function $f(x) = \frac{1}{1+e^{-x}}$. Finally, this result is multiplied by a parameter γ and added to the original score. We obtain by this way the transformed score vector $\widehat{\mathbf{R}} = (r_1, r_2, ..., r_M)$ as it is shown in Equation 5, for $i \in \{1, 2, ...M\}$.

$$v_i \quad \rightarrow \quad \tilde{v}_i = \frac{v_i - \mu_i}{\sigma_i} \tag{4}$$

$$r_i = x_i + \gamma sigmoid(\tilde{v}_i) \tag{5}$$

The parameter γ works as a factor that adjusts the influence of the boost depending on the particularities of the problem. It is optimised using the original scores and contrasting modified scores with expert classifications as ground truth, as it is shown in the experiments in Section 4. Extreme instances with maximum and zero IU ($1.0/M$ and 1.0 scores, respectively) are not modified.

After the transformation is applied, some scores may result in values out of the unit interval, in case their number of votes are located in the right tail of the votes distribution. To avoid this, the transformed score vectors are re-scaled to the interval [0,1] after the transformation using the Equation 6. The x_{min} and x_{max} values are selected amongst the whole set of modified scores.

$$x_{[0,1]} = \frac{x - x_{max}}{x_{max} - x_{min}} + 1.0 \tag{6}$$

The application of this transformation {3} over the example data in Table 1 is illustrated in Table 4. First, we calculate the standard units for the votes in the two classes, taking as example values $\mu_{C1} = 20$ votes, $\mu_{C2} = 30$ votes, $\sigma_{C1} = 3$ votes, and $\sigma_{C2} = 5$ votes. With these values, we take $\gamma = 0.5$ as adjusted for the problem to compute the boost for each example and obtain the modified scores. Since there are a few instances in the data, we omit here the re-scale process after the modification. Once more, we assume two main classes for this example problem.

ID	***N***	***C1***	***C2***	$\boldsymbol{v_1}$	$\boldsymbol{v_2}$	$\boldsymbol{\tilde{v}_1}$	$\boldsymbol{\tilde{v}_2}$	***B. C1***	***B. C2***
0152948451	58	0.414	0.310	24	18	1.33	- 2.40	0.809	0.352
0152863349	14	0.643	0.214	9	3	- 3.67	- 5.40	0.655	0.216
0152878152	33	0.000	1.000	0	33	- 6.67	0.60	0.000	1.000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0152721030	19	0.316	0.263	6	5	- 4.67	- 5.00	0.321	0.266

Table 4: Modified scores obtained from Equation 5 for the two main classes ***C1*** and ***C2*** in the example data presented in Table 1. We consider as example values $\mu_{C1} = 20$ votes, $\mu_{C2} = 30$ votes, $\sigma_{C1} = 3$ votes, and $\sigma_{C2} = 5$ votes, and the value $\gamma = 0.5$. Each row includes the image identification (***ID***), total number of votes (***N***), original scores for the two classes (***C1*** and ***C2***, respectively) original number of votes ($\boldsymbol{v_1}$ and $\boldsymbol{v_2}$), number of votes in standard units ($\boldsymbol{\tilde{v}_1}$ and $\boldsymbol{\tilde{v}_2}$) and boosted scores (***B. C1*** and ***B. C2***). Here we skip the re-scale process due to the reduced number of examples.

3.3. Hybridisation strategy

In this section, we introduce the final procedure leading to the target of the proposed method: exploring the best aggregation of information about the uncertainty in amateur classifications contained in the data. To this aim, we introduce a hybridisation strategy that operates with the three mathematical transformations explained above.

Each transformation tackles one particular expression of the intrinsic uncertainty present in the amateur-labelled data compiled after the project closure. As we discuss in Section 3.1, this uncertainty can be split into three distinguishable forms that are, generally speaking, independent of each other. Within a single instance, the DK votes take part in the distribution of votes across the complete set of classes. However, transformation {1} (Normalisation of the main scores) amends this issue: it restricts the IU to the main classes of the problem, neglecting the influence of other secondary classes; it works isolating the *horizontal* spread of votes across the main classes within a single example. Transformation {2} (DK votes shift), in turn, incorporates the MU codified in

DK votes. Finally, transformation {3} looks at the distribution of votes through the entire population of examples within a same class and boost examples with high confidence. This *vertical* spread is unrelated to the previous two, and adds valuable information about how trustworthy are the main scores for that instance. The information within the data regarding the three uncertainties described here is represented in Figure 2.

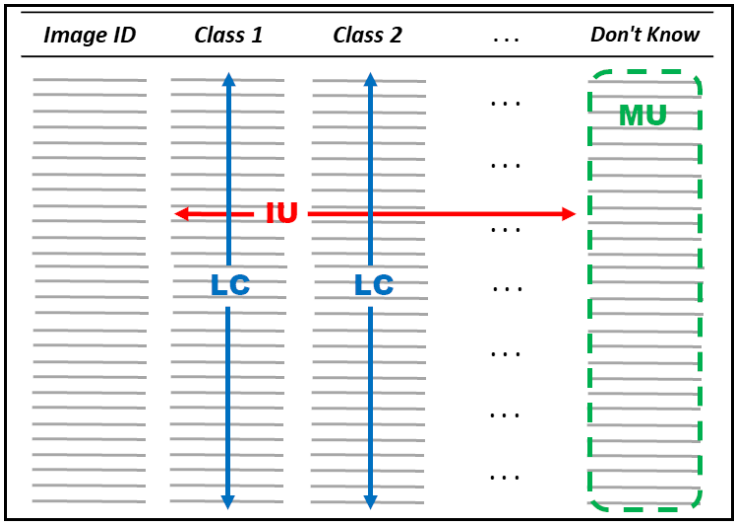


Figure 2: Graphical schema of the three sources of uncertainty present in Citizen Science data.

Hence, a proper blend of the three aggregations is desirable in order to employ the whole information present in the amateur-labelled data. However, the modified scores only focus on one of the uncertainties and ignore the others. Also, depending on the classification problem addressed, these uncertainty types may hold dissimilar relevance within the data. For example, the weight of the DK choice can vary in accordance with the nature of the problem and other factors such as the quality of the images shown to participants. Likewise, when the number of main classes is increased, the distribution of votes may naturally tend to be more uniform. Also, the variability in the LC depends on the running of the project in case a fixed number of votes for each of the instances is required

by project developers.

The method presented here proposes to hybridise the three transformations in all possible combinations to perform a posteriori selection of the best sequence for the problem. The hybridisation performs an independent, sequential, and cumulative application of the single transformations over the reduced score vector, $\widehat{\mathbf{X}}$. As a result, the three transformations are applied following a certain order and taking the modified scores as input of the next transformation. With the three transformations listed above ({1} **Normalisation**, {2} **DK votes shift**, and {3} **Votes boost**), a combinatorial calculation yields we can build a total of $\binom{3}{1} + \binom{3}{2} \cdot 2! + \binom{3}{3} \cdot 3! = 15$ different sequences, which we will denote explicitly from now on by the numerical sequence enclosed by keys[3].

The whole process is developed as depicted in Figure 3. Firstly, taking the amateur-labelled data as input, the method tests all hybrid transformations, where the modified scores work as input of the next transformation of the sequence. A subset of expert classifications allows for the parameters optimisation and for assessing the sequences and ranking them in terms of their quality, using an adequate metric. At the end of the process, the ranking provides a set of improved scores (*Refined Data*) for their later use to obtain final classifications for the objects classified by the crowd of amateurs.

4. Case study: Improving galaxy morphology classification with Citizen Science data

In this section, we illustrate the proposed method with a case study. We look upon the first edition of the Galaxy Zoo (GZ1) project [32], taking the data produced during the run of this project. First, in Section 4.1 we present the particular features of GZ1, concerning the running of the project and the

[3]The factorial terms are a result from the no commutability of the single transformations taken alone. For example, the modified scores generated applying the DK-shift followed by a normalisation ({21} sequence) are not equal if the shift is applied *after* the normalisation ({12} sequence).

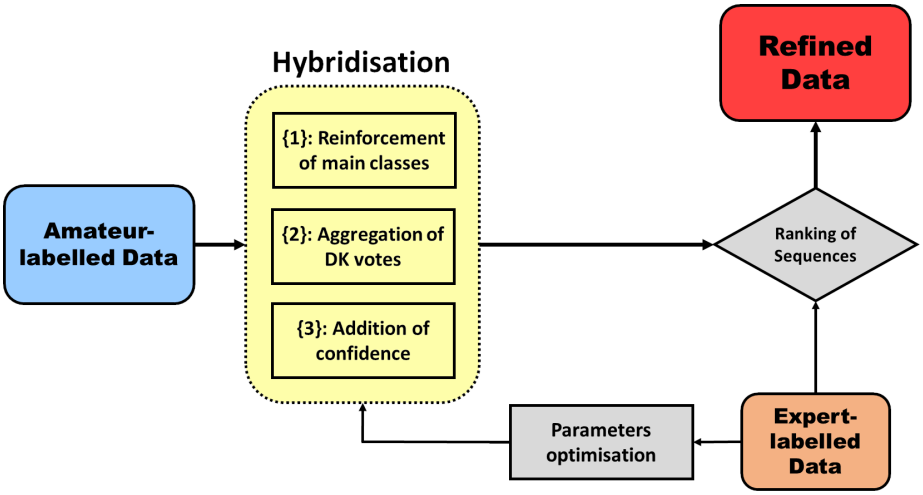


Figure 3: General workflow of the proposed approach. The Amateur-labelled Data obtained after the project closure is enriched by means of a hybridisation of three independent transformations, giving rise to a set of transformation sequences. This process leverages Expert-labelled Data available for the problem, which is used both in the optimisation of parameters and ranking of transformation sequences.

available data. After this, Section 4.2 introduces the two expert catalogues that allow for an assessment of the proposed approach. Then, we describe the experiments implemented for the testing of the method in Section 4.3, and finally we summarise and discuss the results in Section 4.4.

4.1. Galaxy Zoo

The GZ1 project has constituted the very first successful implementation of a Citizen Science project using the Internet. For over a decade it has been bringing together myriads of little efforts from a huge community of amateurs committed to making a contribution to a classical astrophysical problem: the morphological classification of galaxies [24]. This long way has resulted in a list of publications that have supposed a great advances in the astrophysical research [20], via the relaunching of the project in multiple editions as well. Since the first edition of

the project, an application was made available on-line[4], by which any interested individual was able to sign up and start classifying galaxy images from the Sloan Digital Sky Survey[5] (SDSS), one of the main databases of astronomical images compiled to date. GZ1 focused on disentangling the observed bimodality in galaxy morphologies that roughly divides the population between elliptical and spiral galaxies. The first launch caused a great impact, and after six months more than 100,000 volunteers had completed over 40 million classifications for a sample of nearly 900,000 galaxy images [31]. A sample of these images is shown in Figure 4.

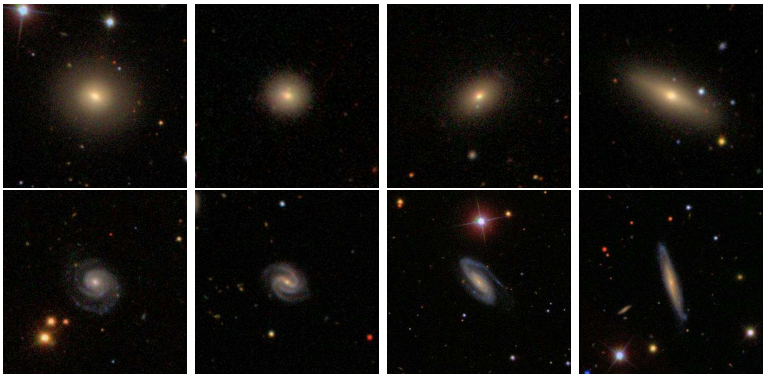


Figure 4: Selection of GZ1 images. The four in the top correspond to elliptical galaxies. The four in the bottom to spiral examples.

In GZ1 project, participants were asked to classify galaxy images choosing between one of six categories: *Elliptical*, *Clockwise Spiral*, *Anti-clockwise Spiral*, *Edge-on Spiral*, *Star / Don't Know*, and *Merger*. Images shown held a common scaling of 423×423 pixels in order to provide a similar basis for all classifications [32]. In this edition, the classification was focused on the distinction between elliptical and spiral morphologies as main classes. However, there are multiple factors that complicate this classification problem. Whereas elliptical galaxies present spherical symmetry, spirals hold plane symmetry. One selection of such

[4]The original GZ1 portal is still maintained at `http://zoo1.galaxyzoo.org`.
[5]`http://www.sdss.org`

images is presented in Figure 4. Consequently, the orientation of the galaxy plays a fundamental role in the identification of its morphology. In addition, the quality of the image strongly depends on several factors such as the distance to the galaxy, and its physical size and brightness. This brings a huge multiplicity of grades of difficulty that is reflected in the uncertainty in amateur classifications.

At the time the project was closed, each image had received an average number of $\sim$38 independent amateur classifications with a standard deviation of $\sim$14 votes, producing the amateur-labelled data of the problem. Then, the GZ1 team started analysing this data to evaluate the influence of biases in the classification task. This resulted in a thorough study by Bamford et al. [4] by which a (manual expert) transformation of the scores obtained from amateurs' votes was developed. Referred as *debiasing* of the scores, it was intended to counter the tendency of classifying blurred images of spiral galaxies as elliptical. As a result, the overall effect was to favour spiral classifications at the expense of elliptical ones. For this amendment, the three spiral sub-categories were joint, giving a combined spiral score (the addition of the *Clockwise*, *Anti-clockwise*, and *Edge-on* scores), which we will refer to as *Spiral* score henceforth.

The GZ1 data was collected in a set of csv files and published[6]. These files include the ID of the galaxy in the SDSS database, the location in the sky, total number of votes received by the galaxy, the set of original scores for all categories, and the debiased scores for the main categories: *Elliptical* and *Spiral*. In addition, the GZ1 team provides final classifications, known as *GZ1 flags*. These are generated via a process that involves the application of a 0.8 threshold over the debiased scores[7]. However, the debiasing of scores required an additional parameter[8] that was not available for the whole GZ1 dataset at the time. Therefore, the debiasing and thus the GZ1 flags were only computed

[6] `http://data.galaxyzoo.org`

[7] Further details about how the GZ1 flags are produced can be found at `http://data.galaxyzoo.org`.

[8] This is the *redshift* of the galaxy, which works as an indicator of the distance to the object.

for a portion of the GZ1 dataset. In the following, we will refer to this sample as *GZ1 subset*, consisting of 667,944 galaxies.

4.2. Expert validation

To validate amateurs' performance through the GZ1, the developers team originally used two expert catalogues [32]. These two expert catalogues will operate as the ground truth needed for the comparison of results. On the one hand, the MOSES catalogue [36] includes 16,516 galaxies present in the GZ1 subset, all of them classified by a team of professional astronomers as *elliptical*. On the other hand, the Longo catalogue [33] includes 25,190 galaxies all labelled as *spiral* by another set of experts and part of the GZ1 subset as well. When both catalogues are compared, we found an overlap of 141 examples, which were removed for the consistency of results. After this adjustment is made, we take the joint expert catalogue, now composed of 41,424 galaxies from the GZ1 subset, which we will refer to as the *validation subset*. This part of the GZ1 data have both expert and amateur classifications. Therefore, it is used to validate the GZ1 flags. Also, as the available expert knowledge on the problem at hand, this subset plays a fundamental role in order to assess the performance of our approach through the following experimental trials. From now on, we will take the validation subset as the ground truth of the problem.

Throughout the set of experiments, we use two metrics for the comparison and validation of results specially convenient for the study of this problem: Accuracy (Acc) and Rejection Rate (RR). As the standard classification measure, the Acc computes the proportion of proper classifications with respect to the number of classified examples. Nonetheless, these classifications are obtained applying a threshold over the scores. In case no score reaches the threshold, the example is annotated as *uncertain*. Hence, the RR measures the fraction of uncertain examples. Taking both measures, we perform a preliminary assessment of the GZ1 flags restricted to the validation subset. This provides a benchmark for the subsequent experiments (Tables 5 and 6, respectively).

This way of validating results involves looking upon GZ1 as a binary clas-

	MOSES	**Longo**	**Joint**
Present in GZ1 subset	16,375	25,049	41,424
Correctly flagged	4,181	20,385	24,566
Incorrectly flagged	1,040	26	1,066
Flagged as *uncertain*	11,154	4,638	15,792

Table 5: Expert validation of GZ flags using MOSES (second column) and Longo (third column) expert catalogues separately, and the joint expert catalogue (fourth column) after removing the 141 overlapped galaxies.

Accuracy	0.9584
Rejection Rate	0.3812

Table 6: Evaluation of GZ1 flags, using the joint expert catalogue over the validation subset.

sification problem. Under this view, *Elliptical* and *Spiral* are the main classes, working as negative and positive classes, respectively, since the identification of spiral patterns entail much more detail and observation. *Merger* and *Star / Don't Know* categories are regarded as secondary classes for which we do not count with any form of expert validation. However, we point to the employment of DK votes to improve the quality of the classifications for the two main classes.

4.3. Experimental setting

Here we present and explain the set of experiments executed for the testing of our approach. In the first place, we illustrate the performance of the three transformations taken independently (Section 4.3.1). After this, we test the hybridisation of the transformations over the GZ1 validation set (Section 4.3.2).

In GZ1 there are two sets of main scores: first, we have the original scores directly obtained from the final count of amateur votes, which we will refer to as *raw* scores. Also, we have the *debiased* scores obtained after the debiasing process explained above. These debiased scores serve of a comparison method

proposed by the experts in [4], as a manual transformation of the raw scores. Here we consider independently both sets of scores for the evaluation of the experiments results.

Similarly to the procedure followed by the GZ1 team, we apply a threshold over the scores in order to assign final classifications to the examples. However, we do not restrict the threshold to one single choice: we explore a series of thresholds in order to get a better intuition about the quality of the classifications provided after certain data transformation has been applied. Here we use six thresholds in the interval [0.5-1.0] taking 0.1 steps, that is, the set (0.5, 0.6, 0.7, 0.8, 0.9, 1.0). These values allow for a well-spread set of cuts that enables a fair comparison between unmodified and transformed scores over a wide range of consensus levels, ranging from 50% consensus (0.5 threshold) to full consensus (1.0 threshold) amongst participants required for the class to be assigned to the object. By this, objects with *Elliptical* or *Spiral* score greater or equal than the threshold being used are labelled as elliptical or spiral, respectively. In any other case, the galaxy is annotated as *uncertain* and counts as not classified. Therefore, each threshold gives one final label for each of the examples in the validation subset, so we can regard each of the thresholds as a single classifier. Likewise, the application of this series over the scores enables us to check the trade-off between Acc and RR as the IU varies across the sample. That is to say: the higher is the threshold, the larger is the amount of *uncertain* galaxies but more accurate the classifications provided.

By using this set of thresholds we compare the quality of the modified scores obtained after applying either a single transformation, or any hybrid combination of them. This is made according to the expert validation explained above. To do this, we represent in a Acc-RR chart the (Acc, RR) points obtained for each of the thresholds in the [0.5-1.0] interval. In addition, for the sake of making the comparison easy and quantitative, along with Acc and RR we consider a third metric: the Hypervolume [50] (HV) subtended by the set of (Acc, RR) points. Since we pursue a two-objective optimisation (we aim to maximise Acc and diminish RR), the HV enables a numerical comparison and ranking of dif-

ferent scores. For its calculation, we take as reference the optimum point (Acc = 1.0, RR = 0.0), that is, the right-bottom corner of the chart. Hence, we are after the minimisation of this measure: the smaller is the HV value, the better is the performance of the transformation.

The last key aspect in the experimental setting is the optimisation of parameters and ranges for transformations {2} and {3} (Equations 2 and 5). A fixed selection of parameters works well for a single application of these transformations. However, the hybridisation of transformations {2} and {3} in distinct order shows that pre-fixed values are not appropriate when we mix these two in variable order. To overcome this issue, we conduct an independent optimisation of the parameters and range of application each time the transformation is applied in a sequence, and therefore using the input scores at the time. This means, for instance, that for sequence {213}, the γ parameter and range of application for transformation {3} is optimised using the scores obtained after application of {2} and {1} in that order, as the sequence establishes in this case. The optimisation is always performed in the same manner:

- First, we select the optimal range of application, squeezing the interval (0.5, 1.0) in 0.01 steps up to a minimum amplitude of 0.2. That is to say, we test the intervals (0.5, 1.0), (0.51, 0.99), (0.52, 0.98),... (0.65, 0.85), and choose that one according to the minimal HV value. In this phase, we use the fixed values $\alpha = 0.1$, $\beta = 1.0$ and $\gamma = 0.4$.

- Secondly, we test the parameters in a range of values. For transformation {2}, parameter α is first tested in the interval [0.01, 1.0] taking 0.01 steps, and then it is done the same with β in the interval [0.1, 10] taking 0.1 steps; for {3}, parameter γ is computed in the interval [0.01, 1.0] taking 0.01 steps as well. For all cases, the value that minimises the HV is selected.

In the hybridisation of transformations, the optimisation of parameters is developed implementing a 70/30 validation. We split the validation subset into two parts: the 70% of the sample is used to perform the parameters optimisation

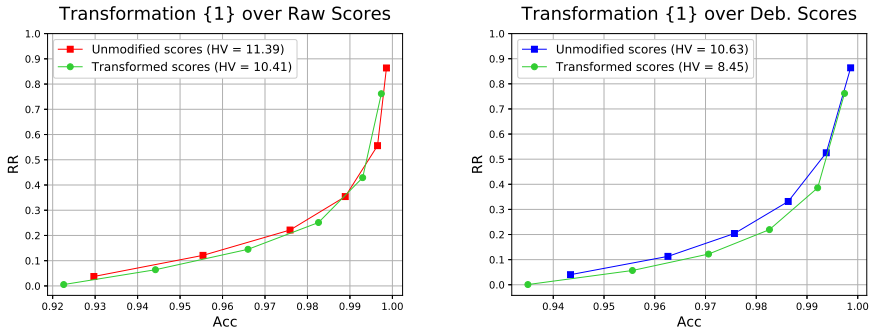
as explained above; the remaining 30% is used for the expert validation of the sequences obtained after completing the parameter optimisation process. This validation ensures that the model is not adjusting the values to the same data used for both optimisation of parameters and evaluation of final classifications.

4.3.1. Single transformations testing

In this first trial of experiments, we test the behaviour of the three proposed transformations (Section 3.2) one at a time, over both raw and debiased scores. Here we adopt a fixed selection of values for the parameters that are found completing a grid optimisation over a reduced range of values. At this point we do not aim to optimise these values, but to use example values to show how the single transformations work.

Transformation {1} develops a normalisation of the main categories scores. Being $M = 2$ for GZ1, with *Elliptical* and *Spiral* the two main categories, we take the reduced score vector $\widehat{\mathbf{X}} = (x_{El}, x_{Sp})$ for each of the examples in the validation subset and calculate the normalised score vector $\widehat{\mathbf{Z}} = (z_{El}, z_{Sp})$ as expressed in Equation 1. For both raw and debiased scores, the application of the set of thresholds brings the charts shown in Figure 5. From left to right, each (Acc, RR) point in the chart corresponds to one of the thresholds in the interval [0.5-1.0], respectively. The HV values indicated in the legend are multiplied by a factor 10^3 for an easier comparison of the quantities.

Transformation {2} introduces a shift into the main scores and uses the DK votes present in the data. To better illustrate the adequacy of this procedure, we firstly check the distribution of this count of votes through the whole population of examples in the GZ1 data. Figure 6 shows the distribution of DK votes across the GZ1 subset. The average number of DK votes is $\mu_{DK} = 2.82$ votes, with a standard deviation of $\sigma_{DK} = 3.55$ votes. The maximum value is $n_{DK_{Max}} = 81$ votes, and there are 153,983 examples ($\sim$23% of the GZ1 subset) for which $n_{DK} = 0$ votes. As before with {1}, we take the reduced score vector $\widehat{\mathbf{X}} = (x_{El}, x_{Sp})$ for each of the examples in the validation subset and compute the shifted score vector $\widehat{\mathbf{W}} = (w_{El}, w_{Sp})$ as indicated in Equations 2 and 3. For



(a) (b)

Figure 5: Test of transformation {1}: the charts show (Acc, RR) points generated by the application of [0.5-1.0] thresholds over raw (a) and debiased (b) scores. Square points represent the unmodified scores, and circular points the scores after the {1} transformation. The HV values indicated in the legend are multiplied by a factor 10^3.

both raw and debiased scores, the application of the set of thresholds brings the charts shown in Figure 7. From left to right, each (Acc, RR) point in the chart corresponds to one of the thresholds in the interval [0.5-1.0], respectively. The HV values indicated in the legend are multiplied by a factor 10^3 for an easier comparison of the quantities. We adopt the values $\alpha = 0.05$ votes and $\beta = 1.0$ votes for calculating the shift (Equation 2), which we select after testing several pairs of values and minimising the HV using expert classifications to compare between original and modified scores. In the same manner, we restrict the range of application to the interval (0.6, 0.9). This means that any object with scores out of this interval is not modified.

Transformation {3} converts the main categories scores, leveraging the distribution of votes in the category through the whole set of instances. This conversion intends to aggregate information codified in the count of votes, so that examples with similar scores but different number of votes can be disentangled for the labelling. As we did with {2}, we first check the distribution of the votes across the population and the two main categories in GZ1. Figure 8 presents the distribution of the total number of votes in the GZ1 subset. The average num-

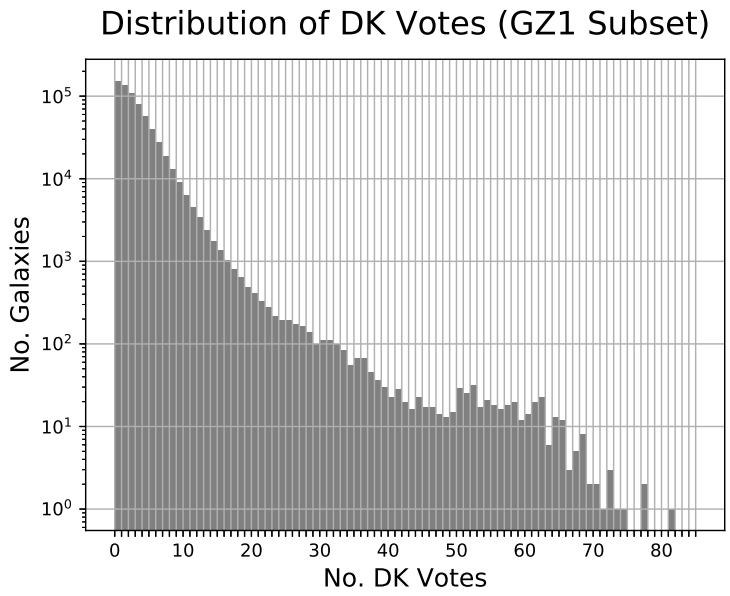
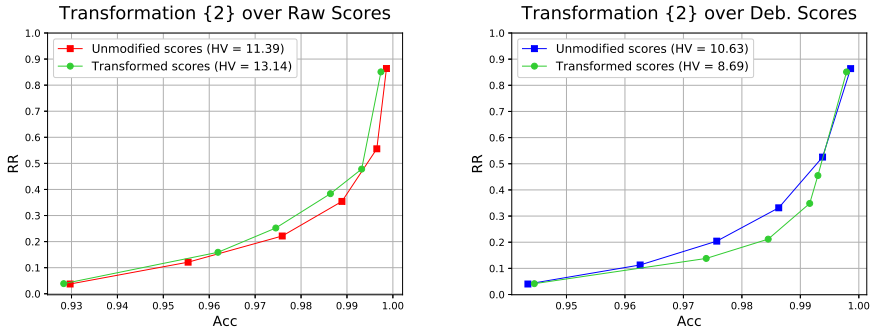


Figure 6: Distribution of DK votes across the GZ1 subset in logarithmic scale. The maximum value is $n_{DK_{Max}} = 81$ votes, with $\mu_{DK} = 2.82$ votes and $\sigma_{DK} = 3.55$ votes.



(a) (b)

Figure 7: Test of transformation {2}: the charts show (Acc, RR) points generated by the application of [0.5-1.0] thresholds over the shifted raw (a) and shifted debiased (b) scores. Square points represent the unmodified scores, and round points the scores after the {2} transformation. The HV values indicated in the legend are multiplied by a factor 10^3.

ber of votes is $\mu_N = 38.76$ votes, with a standard deviation of $\sigma_N = 13.83$ votes. The maximum value encountered is $N_{Max} = 94$ and the minimum $N_{Min} = 4$ votes. For this case, transformation {3} shows a meaningful performance taking as input the normalised scores. Consequently, here we consider the hybrid sequence {13}: we take the normalised score vector $\widehat{\mathbf{Z}} = (z_{El}, z_{Sp})$ for each of the examples in the validation subset and compute the transformed score vector $\widehat{\mathbf{R}} = (r_{El}, r_{Sp})$ as it is shown in Equation 5. For both raw and debiased scores, the use of the same series of thresholds results in the charts shown in Figure 9. From left to right, each (Acc, RR) point in the chart corresponds to one of the thresholds in the interval [0.5-1.0], respectively. As in previous tests, the HV values indicated in the legend are multiplied by a factor 10^3 for an easier number handling. Here, we adopt the value $\gamma = 0.4$ (Equation 5), which we find after optimising the parameter by testing a range of values: we adopt the value that minimises the HV comparing with original scores and using expert classifications as ground truth. Following the same procedure, we also restrict the range of application to the interval (0.6, 0.9).

4.3.2. Hybridisation of transformations

After the testing of the single transformations presented through the previous section, in the following we explain the hybridisation of transformations. In order to extract and combine all information present in GZ1 data, here we propose one hybridisation strategy in two steps: (1) first, we concatenate the three transformations in all their possible combinations; (2) second, we rank the resulting scores according to the HV metric and using the expert classifications. This procedure ensures a proper blend of the transformations in order to aggregate both the information held in DK and total number of votes. We employ the notation $\{xyz\}$ meaning that transformation $\{x\}$ is applied to the scores, then the output scores are used as input to transformation $\{y\}$, and after this, in turn, the result is taken as input to transformation $\{z\}$. When three initial transformations, we can build a total of 15 different sequences: {1}, {2}, {3}, {12}, {13}, {23}, {21}, {31}, {32}, {123}, {132}, {213}, {231}, {312} and

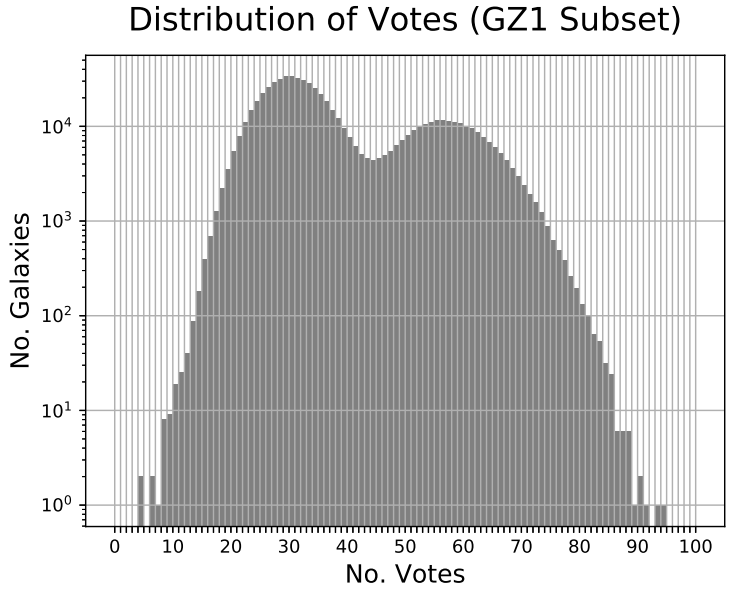
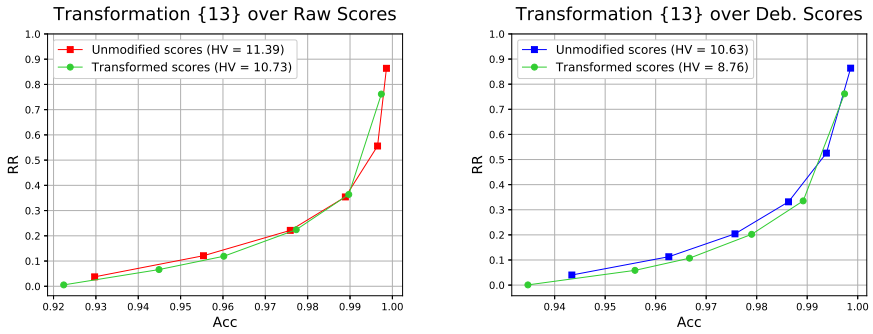


Figure 8: Distribution of total number of votes across the GZ1 subset in logarithmic scale. The maximum value is $N_{Max} = 94$ votes and the minimum $N_{Min} = 4$ votes, with $\mu_N = 38.76$ votes and $\sigma_N = 13.83$ votes.



(a) (b)

Figure 9: Test of sequence {13}: the charts show (Acc, RR) points generated by the application of [0.5-1.0] thresholds over the normalised raw (a) and normalised debiased (b) scores. Square points represent the unmodified scores, and round points the scores after the {13} transformation. The HV values indicated in the legend are multiplied by a factor 10^3.

{321}. In addition, since in GZ1 there are two primary scores, raw and debiased, we take both score types available and compute the whole set of transformation sequences over them. Hence, this hybridisation provides a total of 30 different sets of (Acc, RR) points to compare, after validating each of the final scores obtained with expert classifications (Section 4.2).

As a preliminary trial, we compute and rank the transformation sequences taking the same parameters values used in the previous section. We take $\alpha = 0.05$ votes, $\beta = 1.0$ votes, and $\gamma = 0.4$, and restrict the application range to the interval (0.6, 0.9). Figure 10 shows this ranking of transformation sequences.

Following the parameters optimisation and the 70/30 validation, we complete a second trial computing the same set of hybrid transformations. Figure 11 shows the ranking obtained with this validation for parameters optimisation.

4.4. Discussion of results

We have completed two sets of experiments for the testing of the method. Although the final goal is to obtain the best global transformation to be chosen amongst the set of hybrid sequences for the problem studied, the testing of the transformations alone illustrate how the method works. In broad terms, both experimental trials bring better trade-offs between Acc and RR with respect to the GZ1 benchmark (Table 6). In addition, classifications provided by application of the proposed set of thresholds generally outperform the marks obtained by considering the original scores without modification. In the following, we highlight the most meaningful results in accordance with the experiments presented above:

- In GZ1 we have two sets of scores available, raw and debiased. As it is shown in Figures 5, 7 and 9, debiased scores reach better results compared with raw scores. This trend is maintained in the ranking of transformation sequences (Figures 10 and 11), for which all hybrid transformations applied over debiased scores outweigh those ones obtained from raw scores excepting one: the transformation {3} over debiased scores. These results confirm the critical importance of developing debiasing procedures

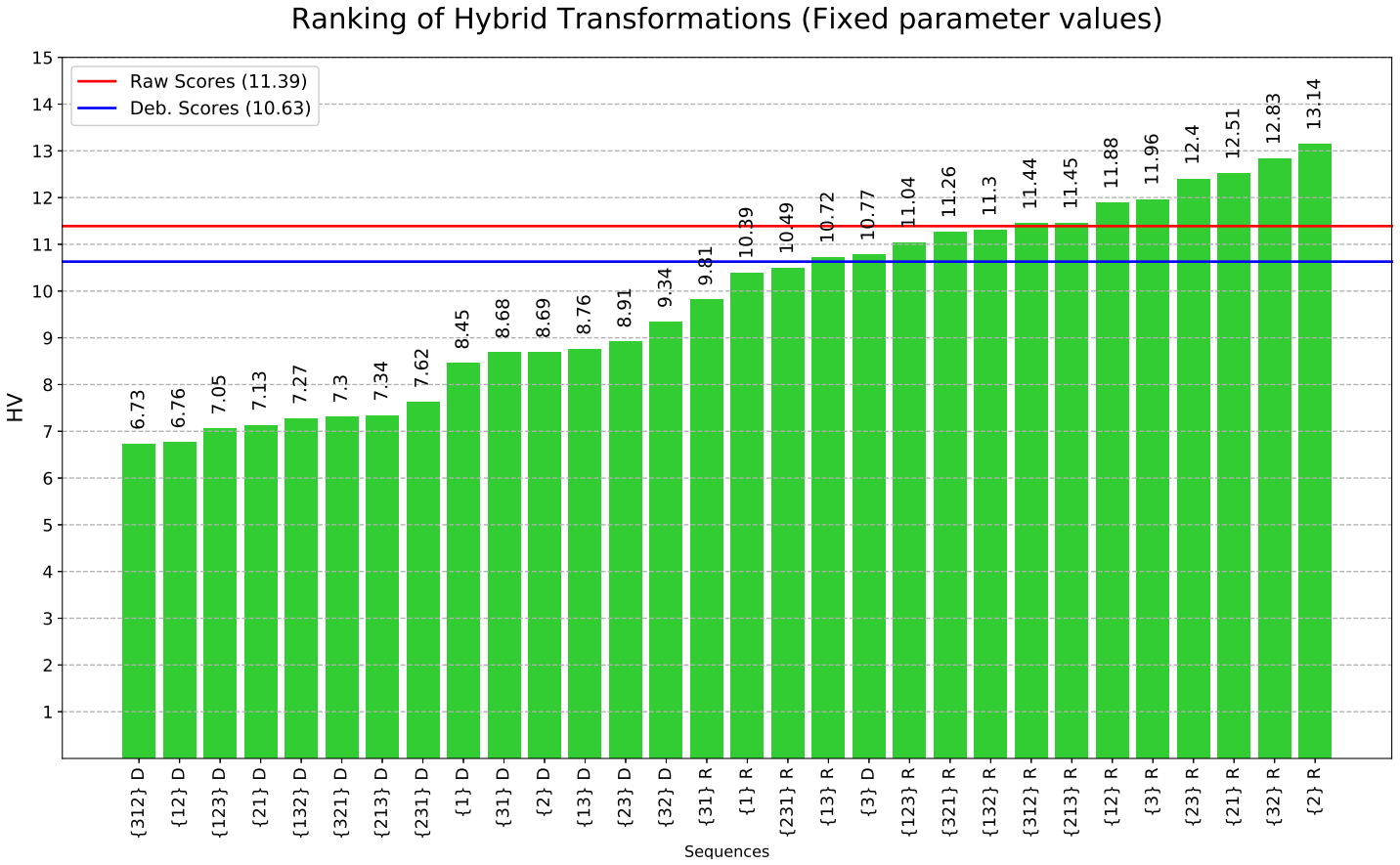


Figure 10: Ranking of hybrid sequences of transformations with fixed values of parameters: $\alpha = 0.05$, $\beta = 1.0$, and $\gamma = 0.4$, and (0.6, 0.9) range of application. The bar height represents the HV value of the sequence indicated in the base, with R standing for Raw scores, and D for Debiased scores. The red horizontal line represents the HV value for initial Raw scores, and the blue horizontal line the same for Debiased scores. All HV values are augmented by a 10^3 factor.

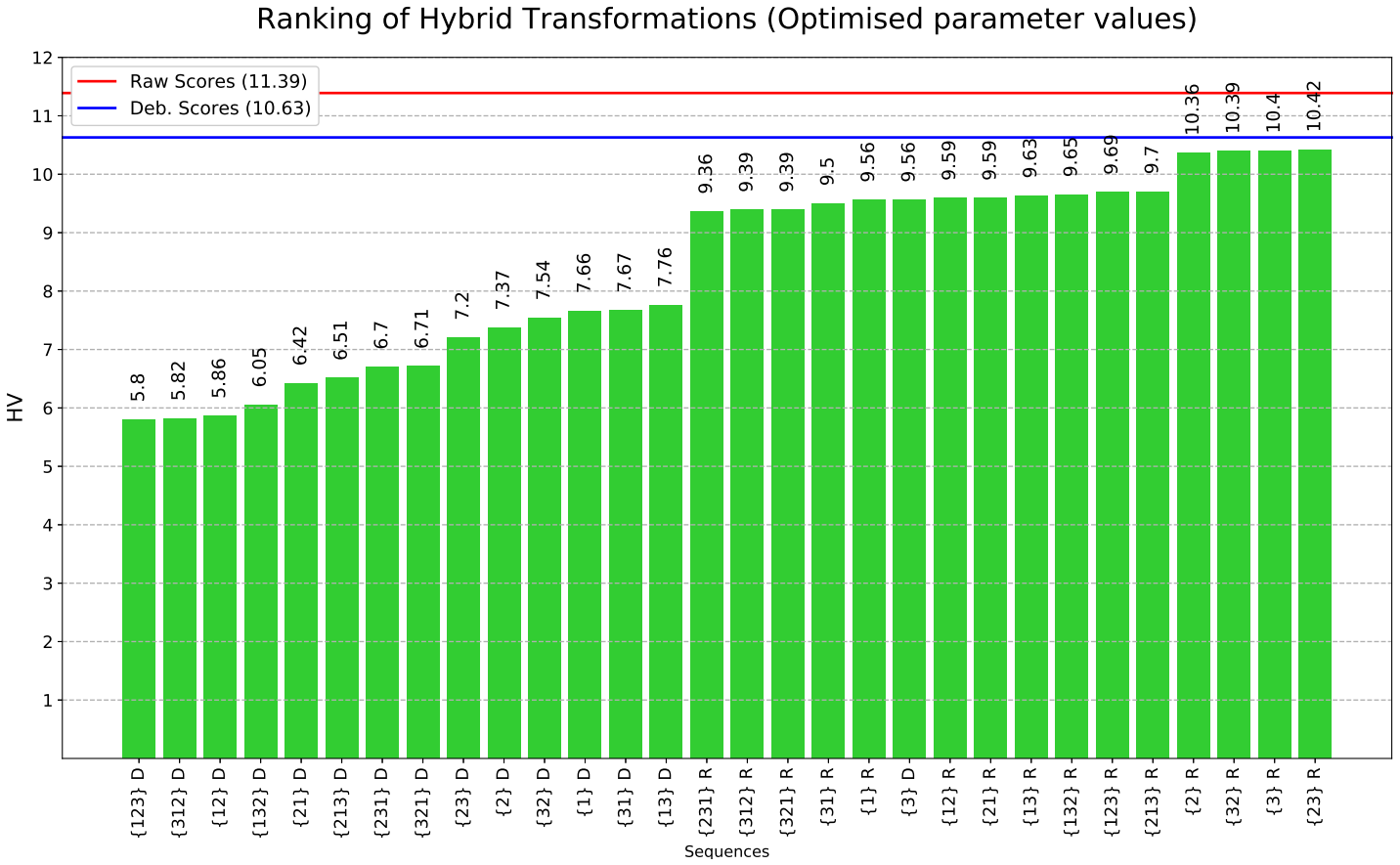


Figure 11: Ranking of hybrid sequences of transformations implementing parameters optimisation and with 70/30 validation of the data. The bar height represents the HV value of the sequence indicated in the base, with R standing for Raw scores, and D for Debiased scores. The red horizontal line represents the HV value for initial Raw scores, and the blue horizontal line the same for Debiased scores. All HV values are augmented by a 10^3 factor.

depending on the problem, which also enables here a comparison between our proposal and a comparable method developed by experts in the field.

- Taken independently, transformation {1} is the only one able to provide a simultaneous improvement of both raw and debiased scores. Transformation {2} worsens the Acc-RR marks for raw scores, and transformation {3} does not provide any improvement to original scores, either raw or debiased. However, through the hybridisation process, it can be seen that {3} is key for the enhancement of the amateur-labelled data: it takes part in 7 of the 10 best sequences for both optimised and not-optimised rankings. This result shows a variety of transformations behaviour, justifying the hybridisation strategy in order to get the most convenient merge of information for the problem being studied.

- The parameters optimisation provides a substantial improvement in the quality of the transformation sequences. Previous to the optimisation, the ranking shows an average HV value of 9.78, with standard deviation of 2.00. After the optimisation, along with the 70/30 validation, the average reduces to 8.37, with standard deviation of 1.58. The best transformation sequence in the optimised ranking, the {123} sequence with debiased scores, gets HV = 5.8, outperforming the best one in the non-optimised ({312} with debiased scores, HV = 6.73) as well as the best result of transformations alone ({3} with debiased scores, HV = 8.69). These results support the adequacy of the optimisation method used. Additionally, all transformation sequences in the optimised ranking outperform the raw and debiased benchmarks, that is, the HV values obtained by unmodified scores (11.39 and 10.63, respectively).

These results confirm the potentiality behind this approach, as able to find an adequate adjustment for the aggregation of information about the uncertainty present in the data, taking the form of either MU or LC, and hidden in the DK votes and distribution of votes through the main categories, respectively. This depends on the choice of metrics for the evaluation of results, and different

metrics could lead to different optimal solutions. However, the results presented here ensure a wide margin of improvement using the proposed method, considering the state-of-art of the problem that is represented by the debiased scores computed by experts.

5. Conclusions and further work

In this paper, we proposed a novel approach for a better employment of the data generated in the course of Citizen Science projects that deal with classification problems. The main achievement of this approach is to be able to aggregate information about different types of uncertainty present in this sort of data: inherent uncertainty, due to the lack of consensus amongst participants that annotate a same example; uncertainty quantified by participants themselves and included as part of the data; and the uncertainty codified in the distribution of votes through the whole dataset for the main classes of the problem. Using this information, our method has proposed three mathematical transformations that modify the original scores and a hybridisation of them that provides the best combined application in accordance with available expert classifications for the problem. To test our approach, we have analysed as case study one of the most representative Citizen Science projects to date, the Galaxy Zoo project. We have presented two sets of experiments: the first one addresses the transformations alone, showing their performance in classifications generated using a threshold over the modified scores; the second implements the hybridisation of the three transformations, demonstrating the advantage of this procedure in order to explore the most adequate blending of them depending on the problem at hand. As a result, the method has proven to enhance classifications accuracy and diminish the amount of unclassified images, comparing with an existing method and using expert classifications as ground truth.

For future work, we plan to extend this approach to more complex settings such as projects involving classification problems with large number of classes, or the aggregation of further information regarding, for instance, participants'

and/or experts' expertise in the classification task. These frameworks will entail new analyses on the aggregation of this sort of data. Eventually, we aim to study the merging of all information available about the problem, pursuing the best results and utility of Citizen Science outcomes for science and research.

Acknowledgement

Funding: The work of M. Jiménez was funded by a Ph.D. scholarship from the School of Computer Science of the University of Nottingham.

References

[1] Alabri, A., Hunter, J., 2010. Enhancing the quality and trust of citizen science data, in: Proceedings of the 6th IEEE International Conference on e-Science, pp. 81–88. doi:10.1109/eScience.2010.33.

[2] Anderson, T., White, S., Davis, B., Erhardt, R., Palmer, M., Swanson, A., Kosmala, M., Packer, C., 2016. The spatial distribution of african savannah herbivores: Species associations and habitat occupancy in a landscape context. Philosophical Transactions of the Royal Society B: Biological Sciences 371. doi:10.1098/rstb.2015.0314.

[3] Ball, N., Brunner, R., 2010. Data mining and machine learning in astronomy. International Journal of Modern Physics D 19, 1049–1106. doi:10.1142/S0218271810017160.

[4] Bamford, S., Nichol, R., Baldry, I., Land, K., Lintott, C., Schawinski, K., Slosar, A., Szalay, A., Thomas, D., Torki, M., Andreescu, D., Edmondson, E., Miller, C., Murray, P., Raddick, M., Vandenberg, J., 2009. Galaxy zoo: The dependence of morphology and colour on environment. Monthly Notices of the Royal Astronomical Society 393, 1324–1352. doi:10.1111/j.1365-2966.2008.14252.x.

[5] Banerji, M., Lahav, O., Lintott, C., Abdalla, F., Schawinski, K., Bamford, S., Andreescu, D., Murray, P., Raddick, M., Slosar, A., Szalay, A., Thomas, D., Vandenberg, J., 2010. Galaxy zoo: Reproducing galaxy morphologies via machine learning. Monthly Notices of the Royal Astronomical Society 406, 342–353. doi:10.1111/j.1365-2966.2010.16713.x.

[6] Barnard, L., Scott, C., Owens, M., Lockwood, M., Tucker-Hood, K., Thomas, S., Crothers, S., Davies, J., Harrison, R., Lintott, C., Simpson, R., O'Donnell, J., Smith, A., Waterson, N., Bamford, S., Romeo, F., Kukula, M., Owens, B., Savani, N., Wilkinson, J., Baeten, E., Poeffel, L., Harder, B., 2014. The solar stormwatch cme catalogue: Results from the first space weather citizen science project. Space Weather 12, 657–674. doi:10.1002/2014SW001119.

[7] Beaumont, C., Goodman, A., Kendrew, S., Williams, J., Simpson, R., 2014. The milky way project: Leveraging citizen science and machine learning to detect interstellar bubbles. Astrophysical Journal Supplement Series 214, 3. doi:10.1088/0067-0049/214/1/3.

[8] Bird, T., Bates, A., Lefcheck, J., Hill, N., Thomson, R., Edgar, G., Stuart-Smith, R., Wotherspoon, S., Krkosek, M., Stuart-Smith, J., Pecl, G., Barrett, N., Frusher, S., 2014. Statistical solutions for error and bias in global citizen science datasets. Biological Conservation 173, 144–154. doi:10.1016/j.biocon.2013.07.037.

[9] Bonney, R., Cooper, C., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K., Shirk, J., 2009. Citizen science: A developing tool for expanding science knowledge and scientific literacy. BioScience 59, 977–984. doi:10.1525/bio.2009.59.11.9.

[10] Bonney, R., Shirk, J., Phillips, T., Wiggins, A., Ballard, H., Miller-Rushing, A., Parrish, J., 2014. Next steps for citizen science. Science 343, 1436–1437. doi:10.1126/science.1251554.

[11] Bonter, D., Cooper, C., 2012. Data validation in citizen science: A case study from project feederwatch. Frontiers in Ecology and the Environment 10, 305–307. doi:10.1890/110273.

[12] Bouveyron, C., Girard, S., 2009. Robust supervised classification with mixture models: Learning from data with uncertain labels. Pattern Recognition 42, 2649–2658. doi:10.1016/j.patcog.2009.03.027.

[13] Brabham, D., 2008. Crowdsourcing as a model for problem solving: An introduction and cases. Convergence 14, 75–90. doi:10.1177/1354856507084420.

[14] Chiclana, F., Herrera-Viedma, E., Herrera, F., Alonso, S., 2007. Some induced ordered weighted averaging operators and their use for solving group decision-making problems based on fuzzy preference relations. European Journal of Operational Research 182, 383–399. doi:10.1016/j.ejor.2006.08.032.

[15] Cohn, J., 2008. Citizen science: Can volunteers do real research? BioScience 58, 192–197. doi:10.1641/B580303.

[16] Crowston, K., Osterlund, C., Lee, T.K., 2017. Blending machine and human learning processes, in: Proceedings of the 50th Hawaii International Conference on System Sciences, pp. 65–73. doi:10.24251/HICSS.2017.009.

[17] Dickinson, J., Zuckerberg, B., Bonter, D., 2010. Citizen science as an ecological research tool: Challenges and benefits. Annual Review of Ecology, Evolution, and Systematics 41, 149–172. doi:10.1146/annurev-ecolsys-102209-144636.

[18] Dieleman, S., Willett, K., Dambre, J., 2015. Rotation-invariant convolutional neural networks for galaxy morphology prediction. Monthly Notices of the Royal Astronomical Society 450, 1441–1459. doi:10.1093/mnras/stv632.

[19] Fedrizzi, M., Pasi, G., 2008. Fuzzy logic approaches to consensus modelling in group decision making. Studies in Computational Intelligence 117, 19–37. doi:10.1007/978-3-540-78308-4_2.

[20] Fortson, L., Masters, K., Nichol, R., Borne, K., Edmondson, E., Lintott, C., Raddick, J., Schawinski, K., Wallin, J., 2012. Galaxy zoo: Morphological classification and citizen science. Machine Learning and Data Mining for Astronomy 11, 118–125. doi:10.1017/S1743921315010911.

[21] Garibaldi, J., Ozen, T., 2007. Uncertain fuzzy reasoning: A case study in modelling expert decision making. IEEE Transactions on Fuzzy Systems 15, 16–30. doi:10.1109/TFUZZ.2006.889755.

[22] Grayson, R., 2016. A life in the trenches? the use of operation war diary and crowdsourcing methods to provide an understanding of the british armys day-to-day life on the western front. British Journal for Military History 2, ISSN: 2057–0422.

[23] Hennon, C., Knapp, K., Schreck, C.J., I., Stevens, S., Kossin, J., Thorne, P., Hennon, P., Kruk, M., Rennie, J., Gada, J.M., Striegl, M., Carley, I., 2015. Cyclone center can citizen scientists improve tropical cyclone intensity records? Bulletin of the American Meteorological Society 96, 591–607. doi:10.1175/BAMS-D-13-00152.1.

[24] Hubble, E., 1926. Extra-galactic nebulae. The Astrophysical Journal 64, 321–373.

[25] Jimenez, M., Triguero, I., John, R., 2018. A first approach for handling uncertainty in citizen science, in: Proceedings of the 2018 IEEE International Conference on Fuzzy Systems (in press).

[26] Kamar, E., Hacker, S., Horvitz, E., 2012. Combining human and machine intelligence in large-scale crowdsourcing, in: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, pp. 467–474.

[27] Klir, G., 1987. Where do we stand on measures of uncertainty, ambiguity, fuzziness, and the like? Fuzzy Sets and Systems 24, 141–160. doi:10.1016/0165-0114(87)90087-X.

[28] Kosmala, M., Wiggins, A., Swanson, A., Simmons, B., 2016. Assessing data quality in citizen science. Frontiers in Ecology and the Environment 14, 551–560. doi:10.1002/fee.1436.

[29] Kuminski, E., George, J., Wallin, J., Shamir, L., 2014. Combining human and machine learning for morphological analysis of galaxy images. Publications of the Astronomical Society of the Pacific 126, 959–967. doi:10.1086/678977.

[30] Li, W., Duan, L., Tsang, I., Xu, D., 2012. Co-labeling: A new multi-view learning approach for ambiguous problems, in: Proceedings of the IEEE International Conference on Data Mining, pp. 419–428. doi:10.1109/ICDM.2012.78.

[31] Lintott, C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., Edmondson, E., Masters, K., Nichol, R., Raddick, M., Szalay, A., Andreescu, D., Murray, P., Vandenberg, J., 2011. Galaxy zoo 1: Data release of morphological classifications for nearly 900 000 galaxies. Monthly Notices of the Royal Astronomical Society 410, 166–178. doi:10.1111/j.1365-2966.2010.17432.x.

[32] Lintott, C., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M., Nichol, R., Szalay, A., Andreescu, D., Murray, P., Vandenberg, J., 2008. Galaxy zoo: Morphologies derived from visual inspection of galaxies from the sloan digital sky survey. Monthly Notices of the Royal Astronomical Society 389, 1179–1189. doi:10.1111/j.1365-2966.2008.13689.x.

[33] Longo, M., 2011. Detection of a dipole in the handedness of spiral galaxies with redshifts z 0.04. Physics Letters, Section B: Nuclear, Elementary

Particle and High-Energy Physics 699, 224–229. doi:10.1016/j.physletb.2011.04.008.

[34] Miller-Rushing, A., Primack, R., Bonney, R., 2012. The history of public participation in ecological research. Frontiers in Ecology and the Environment 10, 285–290. doi:10.1890/110278.

[35] Candido dos Reis, F., Lynn, S., Ali, H., Eccles, D., Hanby, A., Provenzano, E., et al., C., 2015. Crowdsourcing the general public for large scale molecular pathology studies in cancer. EBioMedicine 2, 681–689. doi:10.1016/j.ebiom.2015.05.009.

[36] Schawinski, K., Thomas, D., Sarzi, M., Maraston, C., Kaviraj, S., Joo, S.J., Yi, S., Silk, J., 2007. Observational evidence for agn feedback in early-type galaxies. Monthly Notices of the Royal Astronomical Society 382, 1415–1431. doi:10.1111/j.1365-2966.2007.12487.x.

[37] Schwamb, M., Lintott, C., Fischer, D., Giguere, M., Lynn, S., Smith, A., Brewer, J., Parrish, M., Schawinski, K., Simpson, R., 2012. Planet hunters: Assessing the kepler inventory of short-period planets. Astrophysical Journal 754, 129. doi:10.1088/0004-637X/754/2/129.

[38] Shamir, L., Yerby, C., Simpson, R., Von Benda-Beckmann, A., Tyack, P., Samarra, F., Miller, P., Wallin, J., 2014. Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls. Journal of the Acoustical Society of America 135, 953–962. doi:10.1121/1.4861348.

[39] Show, H., 2015. Rise of the citizen scientist. Nature 524, 265. doi:10.1038/524265a.

[40] Silvertown, J., 2009. A new dawn for citizen science. Trends in Ecology and Evolution 24, 467–471. doi:10.1016/j.tree.2009.03.017.

[41] Simpson, R., Page, K., De Roure, D., 2014. Zooniverse: Observing the world's largest citizen science platform, in: Proceedings of the 23rd Inter-

national Conference on World Wide Web, pp. 1049–1054. doi:10.1145/2567948.2579215.

[42] Sprinks, J., Wardlaw, J., Houghton, R., Bamford, S., Morley, J., 2017. Task workflow design and its impact on performance and volunteers' subjective preference in virtual citizen science. International Journal of Human Computer Studies 104, 50–63. doi:10.1016/j.ijhcs.2017.03.003.

[43] Tsiporkova, E., Boeva, V., 2006. Multi-step ranking of alternatives in a multi-criteria and multi-expert decision making environment. Information Sciences 176, 2673–2697. doi:10.1016/j.ins.2005.11.010.

[44] Wardlaw, J., Sprinks, J., Houghton, R., Muller, J.P., Sidiropoulos, P., Bamford, S., Marsh, S., 2018. Comparing experts and novices in martian surface feature change detection and identification. International Journal of Applied Earth Observation and Geoinformation 64, 354–364. doi:10.1016/j.jag.2017.05.014.

[45] Wright, D.E., Lintott, C.J., Smartt, S.J., Smith, K.W., Fortson, L., Trouille, L., Allen, C.R., Beck, M., Bouslog, M.C., Boyer, A., Chambers, K.C., Flewelling, H., Granger, W., Magnier, E.A., McMaster, A., Miller, G.R.M., O'Donnell, J.E., Simmons, B., Spiers, H., Tonry, J.L., Veldthuis, M., Wainscoat, R.J., Waters, C., Willman, M., Wolfenbarger, Z., Young, D.R., 2017. A transient search using combined human and machine classifications. Monthly Notices of the Royal Astronomical Society 472, 1315–1323. doi:10.1093/mnras/stx1812.

[46] Wu, T., Liu, X., Liu, F., 2018. An interval type-2 fuzzy topsis model for large scale group decision making problems with social network information. Information Sciences 432, 392–410. doi:10.1016/j.ins.2017.12.006.

[47] Yager, R., 2017. Owa aggregation of multi-criteria with mixed uncertain satisfactions. Information Sciences 417, 88–95. doi:10.1016/j.ins.2017.06.037.

[48] Yu, J., Wong, W., Hutchinson, R., 2010. Modeling experts and novices in citizen science data for species distribution modeling, in: Proceedings of the IEEE International Conference on Data Mining, pp. 1157–1162. doi:10.1109/ICDM.2010.103.

[49] Zevin, M., Coughlin, S., Bahaadini, S., Besler, E., Rohani, N., Allen, S., Cabero, M., Crowston, K., Katsaggelos, A., Larson, S., Lee, T., Lintott, C., Littenberg, T., Lundgren, A., Osterlund, C., Smith, J., Trouille, L., Kalogera, V., 2017. Gravity spy: Integrating advanced ligo detector characterization, machine learning, and citizen science. Classical and Quantum Gravity 34, 64003–64025. doi:10.1088/1361-6382/aa5cea.

[50] Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C., Da Fonseca, V., 2003. Performance assessment of multiobjective optimizers: An analysis and review. IEEE Transactions on Evolutionary Computation 7, 117–132. doi:10.1109/TEVC.2003.810758.

Vitae

Manuel Jiménez received his M.Sc. degree in Physics from the University of Granada, Granada, Spain, in 2016. At present, he is a Ph.D. student at the Automated Scheduling, Optimisation and Planning (ASAP) Group, at the School of Computer Science, University of Nottingham. His research addresses the application of machine learning techniques to the classification of astronomical images under extreme conditions of uncertainty.

Isaac Triguero received his M.Sc. and Ph.D. degrees in Computer Science from the University of Granada, Granada, Spain, in 2009 and 2014, respectively. He is currently an Assistant Professor in Data Science at the School of Computer Science of the University of Nottingham. He has published more than 30 international journal papers as well as more than 30 contributions to conferences. He is a Section Editor-in-Chief of the Machine Learning and Knowledge

Extraction journal, and an associate editor of the Big Data and Cognitive Computing journal. He is also a reviewer of more than 30 international journals. He has acted as Program Co-Chair of the IEEE Conference on Smart Data (2016), the IEEE Conference on Big Data Science and Engineering (2017), and the IEEE International Congress on Big Data (2018). He has acted as guest editor for special issues in journals such as Information Sciences, Cognitive Computation, IEEE Access, and Big Data Analytics. His research interests include data mining, data reduction, biometrics, optimization, evolutionary algorithms, semi-supervised learning, bioinformatics and big data learning.

Prof. Robert John received his Ph.D. in type-2 fuzzy systems from De Montfort University, Leicester, UK, in 2000. He joined the University of Nottingham in 2013 as the Head of the Automated Scheduling, Optimisation and Planning (ASAP) Group, School of Computer Science. ASAP has been carrying out highly successful research into the development and application of metaheuristic techniques and hybridisations to scheduling and optimisation problems for the last 12 years. He has authored or co-authored more than 150 publications, an h-index of 32, and more than 5000 citations with papers in the top 1% most cited in ISI. Prof. John is a member of the Editorial Board of the International Journal of Cognitive Neurodynamics, an Associate Editor for the International Journal of Information and Systems Sciences, and an Associate Editor for Soft Computing. He is a member of the EPSRC Peer Review College. A leading researcher in type-2 fuzzy logic. He has held grants that total over 3 million. He currently holds grants worth circa 1 million funded by the UK government and through Horizon 2020. He was the Co-General Chair for the 2007 FUZZ-IEEE International Conference.