Analysis of instance hardness in machine learning using item response theory

Ricardo B. C. Prudêncio[1], José Hernández-Orallo[2], and Adolfo Martínez-Usó[2]

[1] Centro de Informática
Universidade Federal de Pernambuco, Recife (PE) - Brasil
rbcp@cin.ufpe.br
[2] DSIC, Universitat Politècnica de València
Camí de Vera s/n, 46022, València, Spain
{jorallo,admarus}@dsic.upv.es

Abstract. The analysis of instance hardness (or difficulty) can portray valuable insights about the capabilities of machine learning techniques and the demands of benchmark datasets. On the one hand, if we know that a machine learning technique behaves better than another for hard instances, we can prefer that technique whenever we expect a context with more difficult instances. On the other hand, we can evaluate a classifier by the most difficult instances it is able to solve, according to some notion of correctness. This methodology is very similar to the one used in item response theory (IRT), where the ability or proficiency of an individual (e.g., a student) is measured according to the difficulty of the items (e.g., questions) he or she is able to solve. In the current work, we propose to model instance hardness using IRT, which provides a series of tools to characterise instances, including both difficulty and discriminating power. With this characterisation, model assessment in machine learning can also be rethought, by choosing instances that are more discriminative during model evaluation and selection. In this paper, we develop a case study in which instance hardness is measured by fitting the responses of Random Forests with different number of trees. The case study reveals several insights about different levels of discrimination among instances, the adequate number of trees in RF and anomalous situations that were related to noisy instances. Additionally, we explored the idea of producing a model characteristic curve for each classifier, which provides the probability of correct responses given the instance hardness levels. Such curves can be used to select and reuse models for different distributions and levels of instance hardness in a problem.

Keywords: Instance hardness, Item Response Theory, Item Characteristic Curve, operating context, model reuse.

1 Introduction

Hard instances are basically those for which learning models have a low probability of predicting the correct class label. Identifying these instances is in close connection to other hot topics like sample/instance selection, data complexity, noise reduction/filtering and outlier analysis, just to name a few. Indeed, there is a renewed interest in the analysis of instance hardness in machine learning. This research field not only is important to understand the limits of the learning algorithms or to offer information about which instances are misclassified and why [3], but also because it has been recently demonstrated that incorporating instance hardness into the learning process can significantly increase classification performance [11]. In [11] (and in their previous works) Smith et al. present a complete review of these connections and the related applications. They also provide an empirical definition of instance hardness based on the average behaviour of a set of diverse classifiers (e.g., the average error produced by the pool of classifiers for that instance). However, in [11] important information about instance difficulty can be missed by simply averaging the classifiers' outputs for an instance. This is similar to the missed information when a classifier is evaluated by its average performance across the instances in a dataset.

In social sciences, individual traits are also evaluated for a series of tasks. For instance, the verbal ability of a person is a trait that can be evaluated with the use of several exercises that are related to this (latent) trait. In classical test theory (CTT) [9,5], the evaluation approach was very similar to current machine learning practice, that is, given a set of instances (e.g., a dataset), the quality of a model is estimated as the average result for all the elements in the set.

In the area of psychometrics, a new way of evaluating subjects challenged CTT and finally is becoming more and more common in the past decades. This is known as item response theory (IRT) [6,4]. The key issue is that subjects are not evaluated as an average of results for a set of items, but rather as an estimation of the highest difficulty for which items of that difficulty are expected to be solved with a minimum probability of success ν. For instance, if we set this probability to $\nu = 0.75$, if a subject solves all problems of difficulty 1 with 0.95 probability (i.e., solves 95% of them), problems of difficulty 2 with 0.85 probability, problems of difficulty 3 with 0.77 probability, and for every other problem with higher difficulty the probability is lower than 0.75, then we could say that the subject's proficiency or ability is 3. In practice, for robustness, this ability can be estimated using a maximum likelihood estimation (MLE) approach, as we will see.

In order to derive this value of ability, in IRT, item characteristic curves (ICC) are used, which just plot the probability of success on the item (an instance in our case) given the ability of the subject (the performance of a machine learning model). Thus, similarly to IRT, we define the proficiency of a model or technique as the level of hard instances this technique is able to solve. For instance, if a classifier solves all the simple instances but none of the difficult ones, the classifier may be worse (in terms of proficiency) than a classifier that solves most of the simple ones and some of the difficult ones. In other words, instead of just calculating the average accuracy (or other measure) for the given distribution of instances in the dataset, we give more relevance to the difficulty of the instances, slicing the dataset by instance hardness.

This is a very different way of determining whether a technique is better than another. Of course, we expect to see that good learning techniques (in terms of other measures like accuracy) will still have good proficiencies in general. So the analysis will be rather focussed on showing when we can find discrepancies between a 'classical' evaluation (like in CTT) and a more 'modern' evaluation (using IRT). Also, this could be useful to select examples that would be more informative to separate between techniques. For instance, some instances are difficult for some techniques but easy for others, so leading to ICCs that are quite flat, while other instances are difficult for many bad techniques and easy for some very good techniques, having a more segment-like ICC (being more useful instances to *discriminate* techniques).

Apart from ICCs for items we are also interested in analysing how a particular model or technique compares to others for different contexts, where *the context is defined as the expected level of difficulty of the instances*. In this case, the notion of "context plot" defined in [7] is already well defined in IRT as the "person characteristic curve", which is the dual view to the item characteristic curve. In our case, we can talk about a *model characteristic curve*.

In a case study, we adopted Random Forests with different number of trees as the pool of classifiers. The Heart-statlog dataset was adopted to learn and evaluate the classifiers. A 2-parameter IRT model (based on logistic functions) is learned for each instance, fitting the classifiers' correct response probability according to their abilities. We adopt MLE to estimate both the models' parameters of all instances and the classifiers' ability simultaneously. The model parameters characterise both the instance difficulty and its discrimination power. The ability of a classifier is estimated under different instance contexts (levels of instance difficulty) and is apparently related (non-linearly) to its classification accuracy. In addition, ICC plots for detecting anomalous examples are also introduced, showing how in some cases wrongly labelled instances or instances in regions of the instance space dominated by the opposite class can be detected.

Section 2 provides a brief introduction to IRT. Section 3 presents the proposed work, followed by Section 4 in which we present the developed case study. Section 5 presents a discussion with open questions and some future developments. Section 6 concludes the paper.

2 Item Response Theory

Item Response Theory (IRT) [6,4] considers a set of models that relate responses given to items of interest to latent abilities of the respondents. IRT models have been mainly used in educational testing and psychometric evaluation in which examinees' ability is measured using a test with several questions (i.e., items).

In IRT, the probability of a response for an item is a function of the examinee's ability and some item's parameters. There are models developed in IRT for different kinds of response, but we will focus on the dichotomous models. In dichotomous models the response can be either correct or incorrect and we will

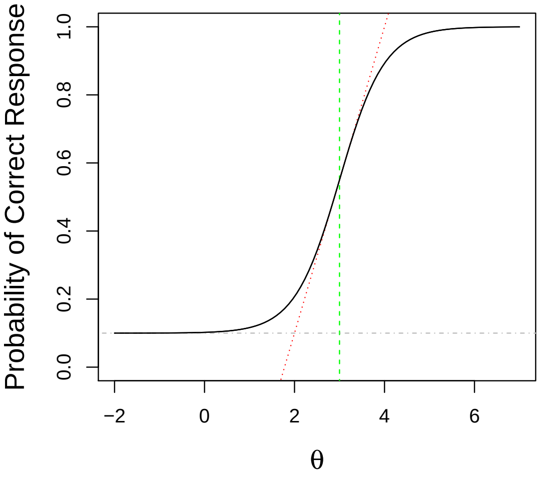


Fig. 1: Example of a 3PL IRT model (in black), with slope $a = 2$ (discrimination, in red), location parameter $b = 3$ (difficulty, in green) and guessing parameter $c = 0.1$ (chance, in grey).

use classification as the easiest machine learning task to show the applicability of IRT in machine learning.

Let U_{ij} be a binary response of a respondent j to item i, in which $U_{ij} = 1$ for a correct response and $U_{ij} = 0$ otherwise. Let θ_j be the ability or proficiency of j. In the basic 3-parameter (3PL) IRT model, the probability of a correct response given the examinee's ability is modelled as a logistic function:

$$P(U_{ij} = 1|\theta_j) = c_i + \frac{1 - c_i}{1 + exp(-a_i(\theta_j - b_i))} \tag{1}$$

The above model provides for each item its *Item Characteristic Curve (ICC)* (see Figure 1 as an example), characterised by the parameters:

- Difficulty (b_i): it is the location parameter of the logistic function and can be seen as a measure of item difficulty. When $c_i = 0$, then $P(U_{ij} = 1|b_i) = 0.5$. It is measured in the same scale of the ability;
- Discrimination (a_i): it indicates the steepness of the function at the location point. For a high value, a small change in ability can result in a big change in the item response. Alternatively we can use the slope at location point, computed as $a_i(1-c_i)/4$ to measure the discrimination value of the instance;

- Guessing (c_i): it represents the probability of a correct response by a respondent with very low ability ($P(U_{ij} = 1| - \infty) = c_i$). This is usually associated to a result given by chance.

The basic IRT model can be simplified to two parameters (e.g., assuming that $c_i = 0$), or just one parameter (assuming $c_i = 0$ and a fixed value of a_i).

The ability of an individual is not measured in terms of number-right answers but it will be estimated based on his/her responses to discriminating items with different levels of difficulty. Respondents who tend to correctly answer the most difficulty items will be assigned to high values of ability. Difficulty items in turn are those ones correctly answered only by the most proficient respondents.

Straightforward methods based on maximum-likelihood estimation can be used to estimate either the item's parameters (when examinees' abilities are known) or the abilities (when items' parameters are known). A more difficult, but common, situation is the estimation when both the items' parameters and respondents' abilities are unknown. In this situation, an iterative two-step method (Birnbaum's method [1]) can be adopted:

- Step (1) Start with initial values for abilities θ_j (e.g., random values or the number of right responses) and estimate the model parameters;
- Step (2) Adopt the estimated parameters in the previous step as known values and estimate the abilities θ_j.

In this method, items' parameters and abilities are simultaneously estimated only based on a set of observed responses to items, with no strong knowledge about the true ability of the respondents.

3 IRT models for instance-wise evaluation

The basic idea of this proposal is to adopt IRT models to evaluate classifiers (or other supervised methods) in the same way that the models are adopted to estimate latent skills of human respondents.

For classification, the value U_{ij} is just understood as a binary response of a classifier (respondent) j to example (item) i. An ICC for an instance (or item) in our context plots the probability of success on the instance, given the ability of the subject (a classifier technique in our case). In other words, instead of computing the average accuracy (or other measure) for the given distribution of instances in the dataset, we estimate classifier ability in terms of instance difficulty. This can be accomplished by adopting the previous methods in IRT to estimate items and classifier parameters when both are unknown.

Given a set of classifiers and a dataset, a general methodology of evaluation can be defined based on IRT models. Initially, a standard validation procedure (e.g., cross-validation) is adopted to collect the binary responses of the classifiers for each instance. Next, the IRT parameters for each instance and the classifiers' abilities are estimated using an iterative MLE procedure (e.g., the Birnbaum's

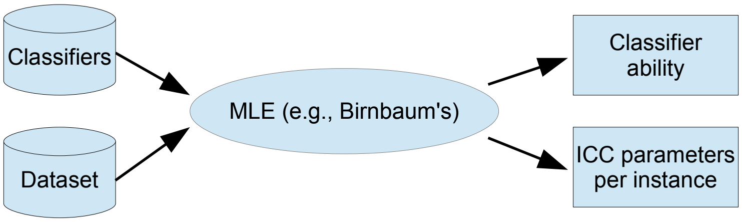


Fig. 2: Evaluation procedure flow chart.

method introduced in Section 2). The IRT parameters and the abilities are estimated to maximise the likelihood of the classifiers' responses observed in the validation procedure. Figure 2 summarises this methodology.

Two kinds of information can be directly acquired from this methodology:

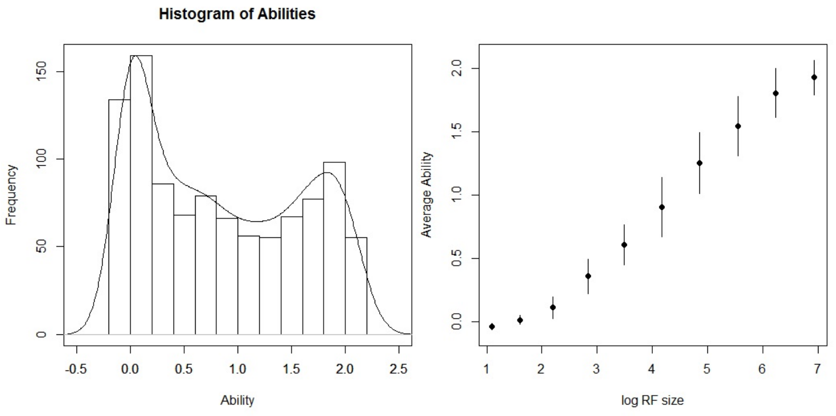
- First the estimated ability is a classification performance measure, which can complement or be compared to other measures like accuracy;
- Second, the ICC parameters can be adopted as a novel way to characterise instances in a dataset. For instance, the difficulty and the guessing parameter can be seen as new instance hardness measures. The discrimination value of an instance in turn can be used to indicate if the instance is useful to distinguish between strong or weak classifiers for a problem.

This is a very different way to characterise instance hardness, instead of considering simple aggregate measures (e.g., average error). Finally, as it will be seen, particular cases (such as mislabelled and noisy instances) can be identified by inspecting the IRT parameters.

4 Case Study

In this case study, a pool of classifiers was produced by Random Forests (RF) trained with different numbers of trees. The values adopted for the number of trees were: $1 + 2^m$, with $m = 1, \ldots, 10$. Theoretically proved in [2] and experimentally confirmed in [8], it is known that both weak and strong classifiers are produced by varying the number of trees, which is convenient for our purposes. By using a limited number of $1 + 2^1 = 3$ trees, the learned classifier would have a low ability to provide correct predictions, while by adopting $1 + 2^{10} = 1025$ trees in turn, we expect to learn a strong classifier. Another alternative is to use a diverse set of heterogeneous classifiers, which we intend to do in future work.

We learned the RFs adopting 10-fold cross-validation, with 100 runs for each parameter setting. Thus each instance was tested by 1000 different RF classifiers (100 runs x 10 parameter settings). For each instance, one IRT model was built using the 1000 binary classifiers' responses for that instance (correct response = 1, incorrect response = 0). In this experiment, we used the Heart-Statlog



(a)

(b)

Fig. 3: (a) Histogram of estimated abilities of RF and corresponding kernel density estimate; (b) RF size (log number of trees) vs. Average Ability. Bars represent the confidence intervals of ability for each RF size.

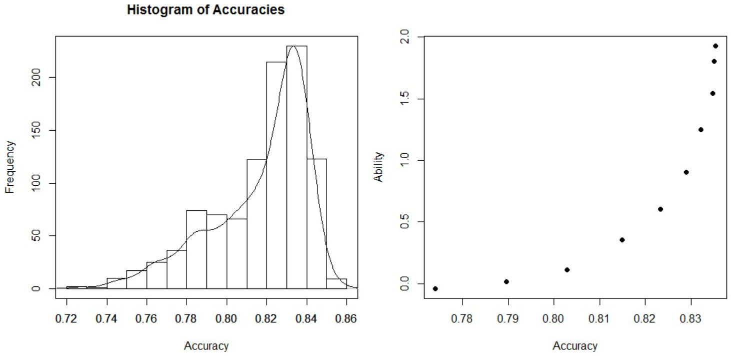
dataset, which has 270 instances. For generating the IRT models, we used the *ltm* R package, which implements the Birnbaum's method [10]. We adopted the 2-parameter model (i.e., no guessing parameter), based on a preliminary battery of experiments. In our experiments, 270 IRT models were built (one per instance) and 1000 values of ability for RF were estimated.

4.1 Classifier Ability

The ability of a classifier is analysed on the basis of whether its responses are correct for instances with different levels of difficulty. As respondents, those classifiers that tend to correctly classify the most difficult instances will be assigned to high values of ability.

Figure 3a presents the histogram and density of abilities for the 1000 respondents (RF classifiers) in the experiments. The plot is bimodal, which suggests the existence of two groups of classifiers separated approximately at ability 1.15. Differences in ability were related to the number of trees adopted in the RF, as expected (see Figure 3b). By splitting the ability at 1.15, we obtain a group of classifiers with a low number of trees, ranging from 3 to 65, and another group of proficient classifiers, with a higher number of trees, ranging from 129 to 1025.

Figure 4a presents the histogram and density estimate for accuracy. Different from ability, the histogram of accuracy does not reveal two distinct groups of classifiers so clearly. Nonetheless, we observe a non-linear monotonic relationship between accuracy and ability (see Figure 4b). Small differences among high accurate classifiers can result in big differences when they are compared in terms



(a) (b)

Fig. 4: (a) Histogram of estimated abilities of RF and corresponding kernel density estimate; (b) Scatter plot showing a non-linear monotonic relation between accuracy and ability of RF with different number of trees.

of ability. This relationship is feasible since instances that were correctly answered by the top accurate classifiers are possibly the most difficult ones. Those instances have more importance in the ability estimation.

4.2 Instance Characteristic Curves

Figure 5 presents the scatter plot of difficulty (b_i) versus discrimination (a_i) of the produced IRT models. The models produced from instances that were *always correctly* classified (39 instances) and instances that were *always incorrectly* classified (1 instance) were excluded in our analysis since they are not useful for discriminating classifiers[3].

ICCs with positive slopes (i.e., positive discrimination values) were obtained for 188 instances (of the 270 total), matching the common assumption of IRT. In these cases, the probability of correct responses is positively related to the estimated ability of the classifiers. At the upper region of Figure 5 we can visualise a cluster of 106 instances (39.2% of the dataset) characterized by high values of the discrimination parameter (greater than 3). A typical ICC in this group is presented in Figure 6a. The instance is correctly predicted by the intermediate and strong classifiers while worse responses are produced by the weakest classifiers. Such instance would be useful to discriminate the weakest classifiers from the rest. Figure 6 presents other examples of ICCs with positive slopes, but with lower discrimination power.

[3] ICCs in these two situations are just lines in the ability range, y-axis equal to 1 for correctly classified instances and y-axis equal to 0 for incorrectly classified instances.

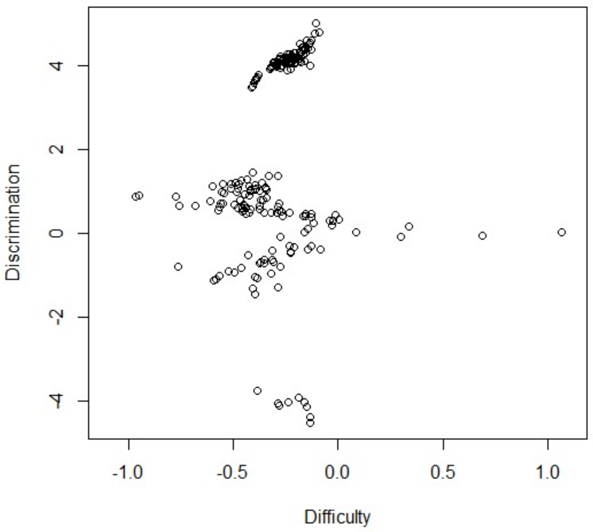
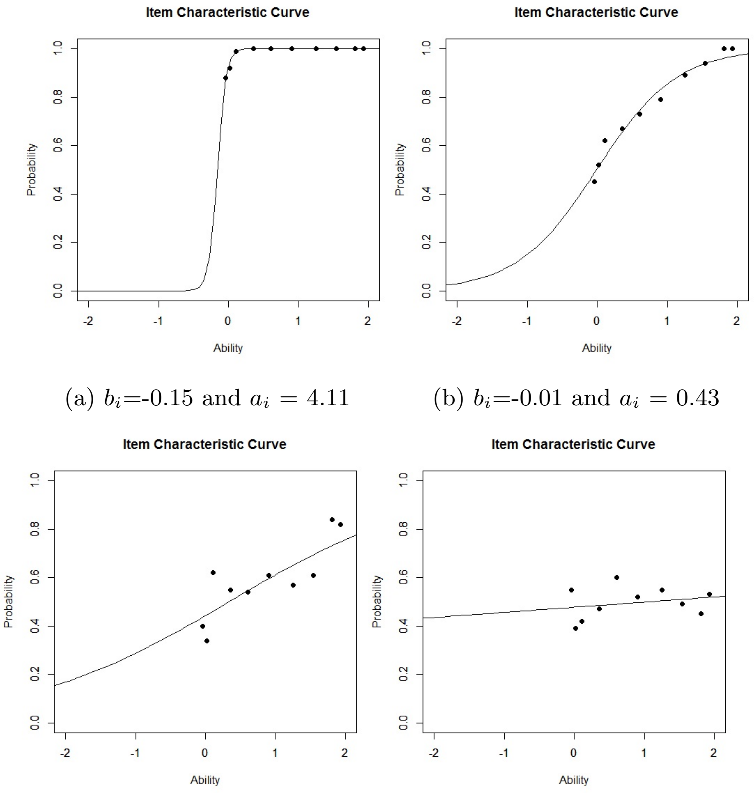


Fig. 5: Difficulty vs. Discrimination scatter plot.

Slightly negative slopes were observed for 42 instances (15.5% of the dataset). In addition, there is a distinct group of 9 instances (bottom part of Figure 5) with high negative values of discrimination (below -3). Figure 7a is a typical ICC in such cases, in which the instances are only correctly classified by the weakest classifiers. Other examples of ICCs with negative slope are presented in Figure 7. These cases are anomalous in IRT (usually referred to as “abstruse” or “idiosyncratic” items) but in our context they may be useful to identify particular situations. For example, if two instances 1 and 2 in a binary classification problem have exactly the same features but belong to different classes, then $P(U_{1j} = 1|\theta_j) = 1 - P(U_{2j} = 1|\theta_j)$. In this situation, one of the instances may have been wrongly labelled, which can result in a negative-slope ICC. Negative slopes also appear for instances that are in regions of the instance space dominated by the opposite class (see Figure 8).

We can relate the IRT parameters to the average error obtained by the classifiers (see Figure 9), which is similar to one of the instance hardness measures proposed in [11]. The discrimination parameter is negatively related to average error (Spearman’s rank correlation = -0.80). The relationship between the difficulty parameter and the average error is not straightforward to see (Spearman’s rank correlation = -0.04) due to the presence of instances with negative slope. When the ICC slope is negative, the difficulty parameter is negatively related to average error (Spearman’s correlation = -0.41), which is not very natural. In



(a) b_i=-0.15 and a_i = 4.11 (b) b_i=-0.01 and a_i = 0.43

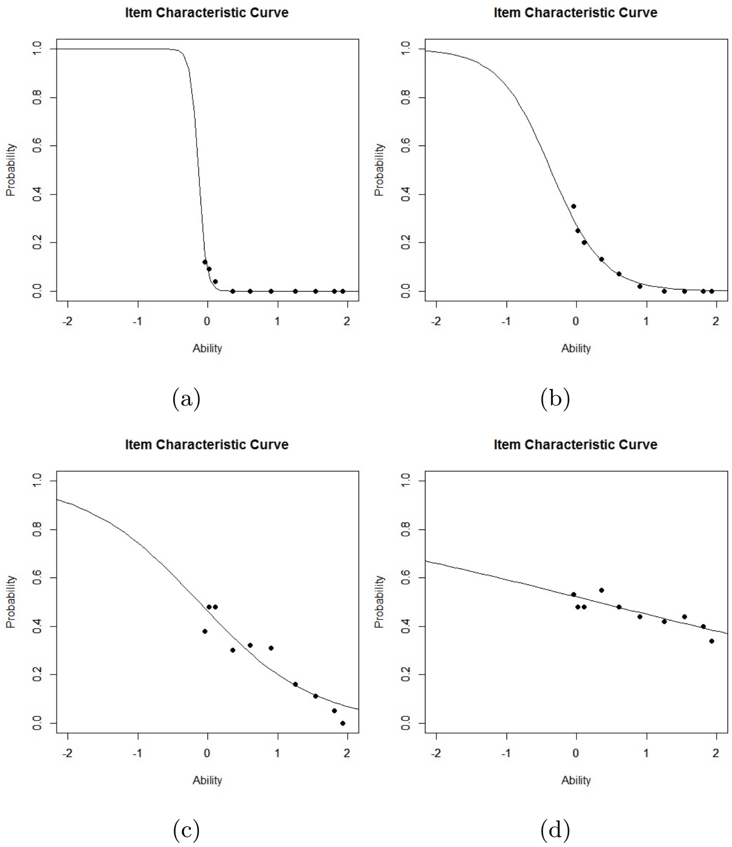
(c) b_i=0.33 and a_i = 0.17 (d) b_i=1.06 and a_i = 0.02

Fig. 6: Examples of ICCs with positive slopes, ranging from high discrimination power (higher values of a_i) to low discrimination power (where parameter a_i is close to zero).

this case, instances are hard by considering average error but easy if we consider directly the difficulty parameter value. As previously said, negative slopes are anomalous cases and have to be treated more carefully.

Based on the above discussion, we further investigate the relation between average error and the IRT parameters, by analysing the product $a_i \times b_i$ (Discrimination × Difficulty). This measure can be directly related to the response probability at the central level of ability $\theta_j = 0$, since in the 2-parameter model we have:

$$P(U_{ij} = 1|0) = \frac{1}{1 + exp(a_i \times b_i)} \tag{2}$$



(a)

(b)

(c)

(d)

Fig. 7: Examples of ICCs with negative slopes with different discrimination power. Such instances represent anomalous cases in IRT are possibly noise or mislabelled instances in the dataset.

For $a_i \times b_i$, we observe an almost monotonic relation to average error, with rank correlation = 0.99 (see Figure 9). This clarifies that average error can be confounding two different things that we see as separate factors in IRT models. In fact, in IRT the usual procedure is to remove the instances with low or negative discrimination, leaving only the instances that are useful to evaluate respondents (i.e., the classifiers) in a meaningful way.

4.3 Model characteristic curves and model reuse

Finally, once the hardness of each instance is estimated, we propose to define a model characteristic curve (MCC) for each classifier of interest, inspired on the concept of person characteristic curve previously developed in IRT. We define in

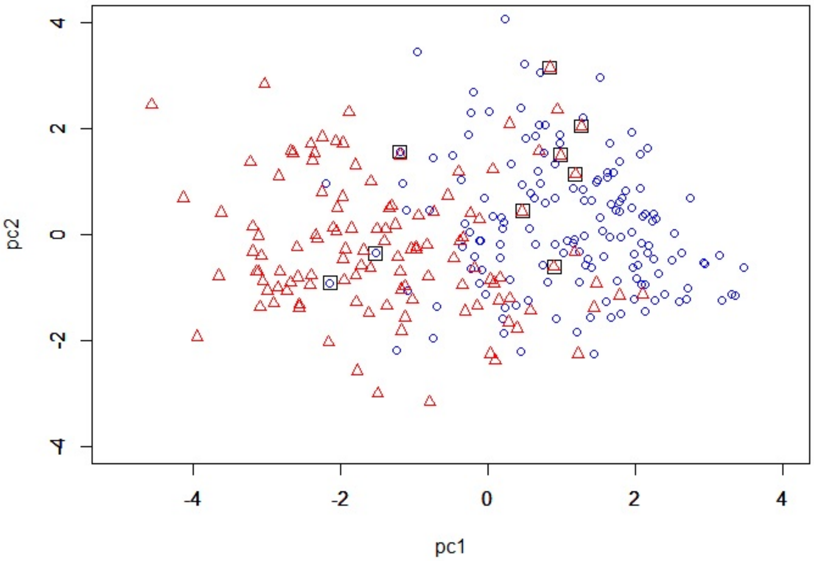


Fig. 8: Visualization of the Heart dataset using the first principal components. Different classes are represented as red triangles and blue circles. Squares indicate the set of 9 instances with the most highly negative slopes.

our work a MCC as a plot for the response probability of a particular classifier along the instance difficulty. Figure 10 presents the MCC of logistic regression (LR), decision trees (DT) and Naive Bayes (NB), for the Heart dataset using the difficulty parameter b_i estimated in previous experiments with Random Forest. Given a classifier (LR, DT or NB), initially, we collected its binary responses for the Heart dataset using cross-validation. The accuracy in this experiment was 0.84, 0.83 and 0.80 respectively for NB, LR and DT. For producing the MCC, we divided the instances in 5 bins ordered by the difficulty parameter[4]. For each bin, we plot in the x-axis the average difficulty of the instances in the bin and in the y-axis we plot the frequency of right responses of the classifier. In this experiment, we excluded the instances with negative slopes, which means that the MCCs were produced using a noise-filtered version of the Heart dataset.

In Figure 10, the three classifiers were equal for the first three bins (easiest instances), corresponding to 38% of the instances considered. From the fourth bin, instances are more difficulty in such a way that it is possible to start distinguishing the classifiers' abilities. In the fourth bin (42% of the instances), DT

[4] Bins were produced in the interval $[0, 1]$ with equal sizes.

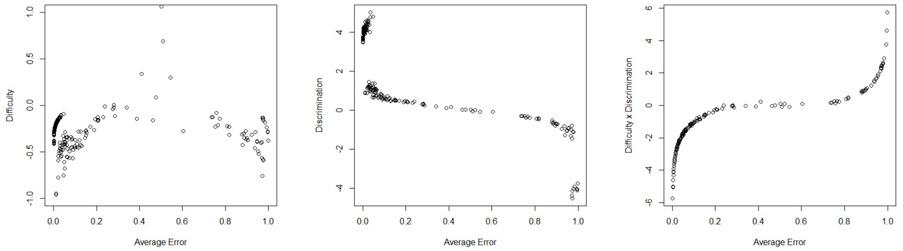


Fig. 9: (a) Average error vs. Difficulty. Spearman's cor. = 0.04. Pearson's cor. = 0.08; (b) Average error vs. Discrimination. Spearman's cor. = -0.80. Pearson's cor. = -0.79; (c) Average error vs. (Discrimination × Difficulty). Spearman's cor. = 0.99. Pearson's cor. = 0.83.

obtained the worst response probability (0.9), while the response probability of LR was slightly higher than NB (0.97 against 0.96). Finally, for the fifth bin (20% of the instances), NB was the best classifier, followed by LR and DT. The response probability was 0.86, 0.81 and 0.76 respectively for NB, LR and DT.

We could imagine some scenarios in the Heart problem where the MCCs would be adopted for selecting different algorithms. For instance, if a higher number of easy instances is expected during a deployment context, then the MCCs would indicate that there is no advantage among the three classifiers in terms of their correct responses. Hence, one could choose any classifier based on particular preferences (e.g., one could prefer to a DT because it is easy to interpret). For the intermediate difficulty (fourth bin in Figure 10), LR and NB would be indicated. Finally, the advantage of NB on the Heart problem would be higher for contexts when it is expected a higher frequency of more difficulty instances. To make such decisions, one has to define the difficulty of instances during deployment. This can be done by defining a new classification task where the predictor attributes are the original attributes of the problem and the class label is the category of difficulty defined by analysing the MCCs (see Figure 11).

5 Discussion

In this paper, we provide a completely different way of analysing classifiers at instance level. The notion of instance hardness we are using is population-dependent (it depends on the set of techniques used) as in previous work (e.g.,[11]). Nevertheless, no previous work explores the variance of the classifier responses or tries to fit models. Additionally we provide a different way of deciding whether one technique is better than another, relating the probability of correct response to the level of instance hardness (the MCCs). Although the experiments are not conclusive in many aspects (since it is a case study on a single dataset and pool of classifiers), some interesting insights were revealed. There is a number of different issues that we can investigate in future work:

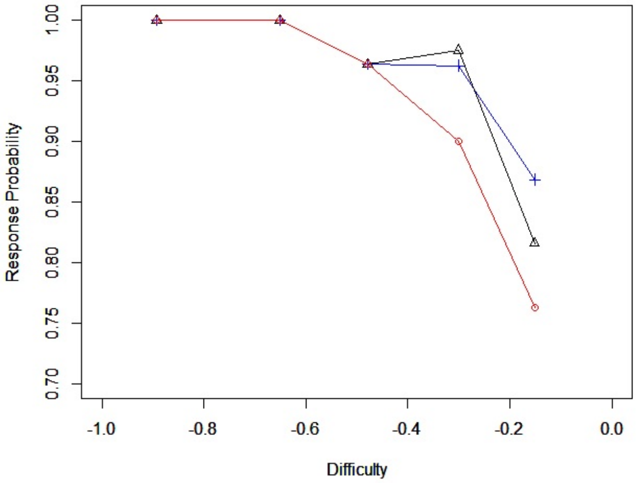


Fig. 10: MCC plots for the NB, LR and DT classifiers considering the instances with positive slopes (i.e., relevant instances) in the Heart dataset. Legend: NB - blue crosses; LR - black triangles; DT - red circles.

- **Alternative IRT models and classifiers**: in our case study we applied the 2-parameter logistic function as IRT model. This was supported by a preliminary battery of experiments in which no significant gain in model fitting was observed by adding the guessing parameter. In fact, when we tried the 3-parameter model the values of the guessing parameter were in general very close to zero. Actually the outputs of the RFs are deterministic, i.e., there is no guess (we adopted odd numbers of trees and we solved a binary classification problem). For different pools of classifiers, however, the guessing parameter can be important and it will be probably related to the class distribution. In our work, we only adopted dichotomous IRT models since we tried to model binary responses of classifiers. However, learning models can return scores and probabilities as well. Similarly, regression models can be considered. In order to model such responses, other IRT models from the literature will have to be adopted. Finally, it is not clear that instance hardness measured using RFs can be extrapolated to evaluate other types of classifiers, as we performed in the previous section by inspecting the MCCs for LR and DT. A possibility is to use a pool of diverse and heterogeneous classifiers as in [11];
- **IRT models for evaluation**: once we have ICC for all examples of a dataset using a pool of techniques, we can derive the proficiency of a new technique with very few instances. Instead of using cross-validation we could just use

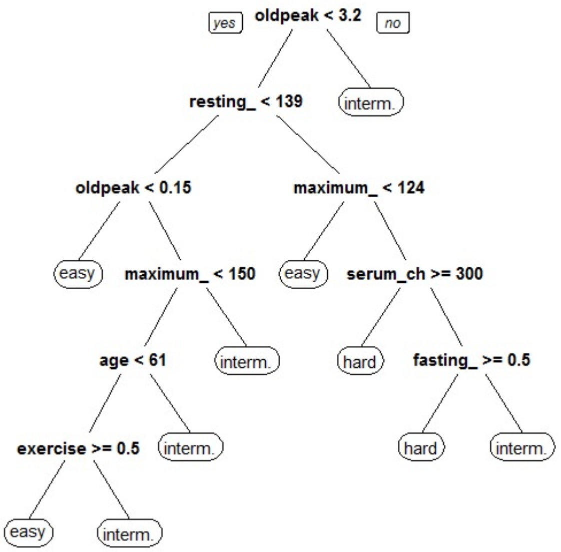


Fig. 11: Decision tree built to predict the category of difficulty in the Heart problem. Labels were defined according to the MCCs (label *easy*: three first bins in Figure 10; *interm.*: fourth bin; *hard*: fifth bin).

a subset with a few discriminative examples as hold-out, and train with all the rest. With this we could train with almost all examples, and test with very few examples and infer a reasonably good estimation of proficiency using MLE. Considering experiments across different datasets, a notion of proficiency or ability per classifier and dataset would make aggregations commensurate if we used different datasets. This would solve one common problem in current machine learning experimentation, as it does not make sense to make the average of non-commensurate metrics, such as accuracy, of several classifiers. Here, if the proficiencies are normalised, we have that items from *different* datasets can be compared in terms of difficulty;

- **IRT models for learning**: the concepts of instance difficulty and discrimination can be adopted in iterative learning procedures like boosting. For instance, the most informative items to select in boosting could be the ones with high expected variance given the inferred proficiency of the classifier so far, i.e., the instances with highest slopes for the proficiency of the classifier;
- **IRT models for instance filtering**: a straightforward application of instance hardness is to filter out the most difficulty instances in a dataset, aiming to produce a better dataset for training, as done in [11]. A similar

idea can be done using IRT models, but now considering both difficulty and discrimination. Negative slopes can be informative in this issue;

- **IRT models for context change**: we also ask the question of how a model can be reframed to give more relevance to hard instances over easy instances, by a cost-sensitive modification of the performance metrics, so that models can be reused for different distributions and levels of instance hardness.

6 Conclusion

In this paper we proposed the use of IRT for an instance-wise evaluation of classifiers. As previously discussed, there could be many applications of this psychometric procedure for machine learning. Of course, there might be criticisms too. For instance, the IRT approach may not be the best way of looking at comparing classifiers or techniques in front of items of varying difficulty. Also, it can be computationally expensive to fit IRT models for millions of instances, which is very common now in real applications. We expect that this paper can result in interesting discussions and also questions to be answered in the future.

References

1. A. Birnbaum. *Statistical Theories of Mental Test Scores*, chapter Some Latent Trait Models and Their Use in Inferring an Examinees Ability. Addison-Wesley, Reading, MA., 1968. 5
2. Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 6
3. Carlos Eduardo Castor de Melo and Ricardo Bastos Cavalcante Prudêncio. Cost-sensitive measures of algorithm similarity for meta-learning. In *Intelligent Systems (BRACIS), 2014 Brazilian Conference on*, pages 7–12, Oct 2014. 2
4. Rafael Jaime De Ayala. *Theory and practice of item response theory.* Guilford Publications, 2009. 2, 3
5. Steven M Downing and Thomas M Haladyna. *Handbook of test development.* Lawrence Erlbaum Associates Publishers, 2006. 2
6. S. E. Embretson and S. P. Reise. *Item response theory for psychologists.* L. Erlbaum, 2000. 2, 3
7. José Hernández-Orallo, Ricardo Bastos Cavalcante Prudêncio, Meelis Kull, Peter Flach, Ahmed Chowdhury Farhan, Nicolas Lachiche, and Adolfo Martínez-Usó. Reframing in context: A methodology for model reuse in machine learning. *Artificial Intelligence Communications*, Under review. 3
8. Patrice Latinne, Olivier Debeir, and Christine Decaestecker. Limiting the number of trees in random forests. In *In Proceedings of MCS 2001, LNCS 2096, 2001*, pages 178–187, 2001. 6
9. Samuel A Livingston. Item analysis. pages 421–441. Lawrence Erlbaum Associates Mahwah, 2006. 2
10. Dimitris Rizopoulos. ltm: An R package for latent variable modeling and item response analysis. *Journal of Statistical Software*, 17(5):1–25, 2006. 7
11. Michael R Smith, Tony Martinez, and Christophe Giraud-Carrier. An instance level analysis of data complexity. *Machine learning*, 95(2):225–256, 2014. 2, 9, 13, 14, 15