Comparing supervised machine learning algorithms using item response theory models

Mariana Curi* (presenting author)
Universidade de São Paulo, São Carlos, Brazil - mcuri at icmc dot usp dot br

Arnaldo Candido Junior
Universidade Tecnológica Federal do Paraná, Medianeira, Brazil - arnaldoc at utfpr dot edu dot br

Nathan Siegle Hartmann
Universidade de São Paulo, São Carlos, Brazil - nathansh at icmc dot usp dot br

Abstract

Many supervised learning algorithms can be adopted to automatically classify texts into different levels of complexity in order to assist educators in classroom routine. Traditionally, the comparisons of these classifiers are based on simple measures like accuracy, Kappa coefficient or ROC curve, for instance. The present work applies Item Response Theory (IRT) models to compare classifiers, contributing for a better understanding of the results of machine learning experiments. Four IRT models were considered: two and three parameter logistic models, graded response model and longitudinal IRT model. Preliminary results show that item parameters can be used to identify instances (texts) with noise or some particular inconsistency, or whether the classifiers overfit. It is possible to rank classifiers according not only to their accuracy, but also taking into account the characteristics of the instances that are correctly or incorrectly classified. An important and useful advance was made fully understanding IRT parameter meaning in the context of classification problems.
Keywords: artificial intelligence; latent trait analysis; item response theory; supervised learning.

1. Introduction

The Programme for International Student Assessment (PISA)[1] provides education rankings based on international tests taken by 15-year-olds in maths, reading and science. The tests, run by the Organization for Cooperation and Economic Development (OECD) and taken every three years, have become increasingly influential on politicians who see their countries and their policies being measured against these global school league tables. Asian countries dominate the ranking of reading skills. Singapore, China and Canada are the countries with best reading skills. Other countries as Germany, USA and UK do not even figure in top 10. The OECD reported rated UK teenagers aged 16 to 19 the worst of 23 developed nations in literacy and 22nd of 23 in numeracy. It also showed results for Brazil below the average of the countries surveyed. 56.6% of Brazilian students did not reach the levels considered minimum in reading, which means that, at best, they can only recognize themes of simple and familiar texts. Furthermore, only 8.3% of Brazilian students reached maximum reading levels, being able to deal with complex texts and perform in-depth analysis on such texts.

The development of reading skills has long been related to success in future academic and professional activities. Aimed at raising the quality of the teaching model for reading and text comprehension in this country and trying to close some gaps in Brazilian public policies for education, many features and computer systems for the Brazilian Portuguese have been launched recently. An example is the First Book Project (*Projeto Primeiro Livro*)[2], which helps children and young people from public schools to learn grammar, spelling and develop narratives. Another example is the Victor Civita Foundation, sponsored by the publishing house Abril, which supports teachers, school managers and public policy makers of Elementary Education with

[1] Available at ⟨http://www.oecd.org/pisa/⟩.
[2] Available at ⟨http://www.primeiro-livro.com⟩

lesson plan search engines, social network for educators to exchange experience and share knowledge, and a resource bank for classes[3].

Currently, in Brazil, the elementary school is divided into two stages - 1st to 5th year, and 6th to 9th year. The National Curriculum Parameters (1998), however, divide these two stages into four cycles. In this article, we focus on three stages: the end of the first cycle (3rd year), the second (4th and 5th years) and third cycles (6th and 7th years), because they are fundamental for students to achieve adult reading comprehension.

There are some tools for Brazilian Portuguese such as the Flesch Index (Martins et al., 1996), which is adapted for Portuguese and used in the Microsoft Word, and mainly the Coh-Metrix-Port and AIC, developed in the PorSimples project (Aluísio and Gasperin, 2010), whose goal is to simplify Web texts for people with poor literacy levels. These tools, however, do not meet the needs of educators in the classroom: there are no classifiers able to discriminate the level of complexity of each year focus of this study 3rd to 7th years, using metrics of the many language levels.

For the English language, there are tools for classifying reading materials for children used in US schools, based on both quantitative data such as Lexile[4] (Stenner, 1996, Lennon and Burdick, 2004) and better informed such as Text Easability Assessor (TEA)[5] that uses Coh-Metrix (Graesser et al., 2004, Graesser et al., 2011) metrics.

Automatic text classification and recommendation helps in efforts to mitigate the presented literacy problems. It offers teachers a way to select suitable texts to their students. Ideally, in order to build such a classifier, one should select a good learning algorithm and adjusts its parameters in the best way possible.

Classifiers are traditionally evaluated and compared by several measures like accuracy, F-measure, Kappa, Receiver Operating Characteristic (ROC) Area, among others (Witten and Frank, 2005). However, there is little research on Item Response Theory (IRT) for the task and some challenges should be overcome for the meaningful interpretation of IRT results in the comparisons of classifiers. Taking inspiration from the proposal of Martínez-Plumed et al., 2016, we apply four different IRT models for comparing classifiers, contributing for a better understanding of the results of machine learning experiments. Each text (instance) is considered as an item and each classifier as an subject. Besides classifier comparison, the approach also allows us to evaluate instance in the sense of which instances are best to train classifiers, which ones are on decision boundaries in instance space and also to detect outliers and noise data.

2. Methods

The methods applied in this work consisted in training and testing several classifiers, storing their predictions to each test instance and then using the results as input to IRT models.

2.1 Experiments

In order to run the experiments, a dataset with features extracted of 1.448 texts was used to train and to test several classifiers. Each text was manually annotated by experts in the area. These texts were labeled as fitted for one of three bands of text complexity aimed on this work: the 1st, 2nd or 3th cycle of Elementary School. Overall, 188 features were extracted from each text. The features were based on the Coh-metrix research with several additions (Graesser et al., 2011). To avoid feature domimance, all feature values were normalized to have 0 mean and standard deviation 1.

The extracted data was then used to train 96 classifiers from 20 different learning algorithms available on Weka[6] machine learning suit (Hall et al., 2009). Classifiers are analyzed according to their ability to correctly identify a text recommended cycle. The difference between classifiers obtained from the same learning algorithm are the parameters used for training. To generate the models, 10-fold cross-validation was applied. Each classifier was executed twice, first with the original dataset and then with a modified version of the dataset generated by the PCA method (Principal Component Analysis) aiming at dimensionality reduction and covering 90% of data variance. Thus, 188 experiments were performed, coincidentally, the same number of features.

[3]Available at ⟨http://www.rede.novaescolaclube.org.br⟩

[4]Available at ⟨lexile.com⟩

[5]Available at ⟨tea.cohmetrix.com⟩

[6]Available at ⟨www.cs.waikato.ac.nz/ml/weka/⟩.

Table 1: Overview of the performed experiment.

Algorithm Family	Algorithm	Tested Parameters	Models per Dataset
Bayes	BayesNet	estimators (4), search methods (3)	5
	NaiveBayes	kernel estimator (2), discretization (2)	3
Decision Tree	J48	confidence factor (3), minimum objects (2), prunning (2)	4
	LMT	default (1)	1
	RandomForest	max depth (3)	3
	RandomTree	max depth (3)	3
	REPTree	default (1)	1
Ensemble	Stacking	SMO over three other algorithms	1
	Vote	combination rules (3) over 3 algorithms	3
K-Nearest Neighbours	IBK	neighbours (5), distances (2), weighting (2)	18
	KStar	default (1)	1
	LWL	neighbours (6), distances (2), classifiers (3)	9
Logistic Regression	Logistic	gradients (2)	2
	SimpleLogistic	default (1)	1
Neural Network	MultiLayerPerceptron	decay (2), momentum (5), epochs (2), learning rates (5), layers(3), total neurons (7)	21
Rule based	DecisionTable	neighbours (2)	2
	JRip	prunning (2)	2
	OneR	default (1)	1
	PART	confidence factor (3), minimum objects (2)	3
Support Vector Machines	SMO	margin complexity (5), kernels (4), polynomials (5)	10

The best experiment presented results comparable to state-of-the art classifiers for English (Feng et al., 2010) (despite they aimed on 4 levels of difficulty), achieving 76.93% accuracy and 78.1% F-measure. Table 1 presents an overview of the classifiers used in each experiment (first column), the parameters analyzed and the tested values for each (second column), and the total of induced models in each dataset (third column). There are 94 models for each dataset, totalizing 188 experiments.

The chosen parameters are the ones most likely to affect the classification results. Some parameters, considered less important in this process, as random seeds, were ignored. Additionally, some classifiers with few parameters like Simple Logistic were tested only once in each dataset. Classifiers with multiple experiment contais one experiment using default settings. The other experiments differs minimally from default settings, changing few parameters at time. This is done partially because some parameters are highly dependent among themselves (eg. the kernel and the polynomial degree in SMO) and partially to avoid an explosive number of experiments (eg., number of neurons, epochs and learning rates in neural networks). An exception are the K-nearest neighbors algorithms, which distances measures and number of neighbors are combined in several experiments. Most parameters with two values tested are binary, for example, the use of pruning in decision trees.

2.2 IRT models

We considered four different IRT model: the two (2-PL) and three (3-PL) parameter logistic models (Birnbaum, 1968), the graded response model (Samejima, 1969) and an extension of the multidimensional Rasch model for the repeated administration of the same items to a sample over different occasions proposed by Andrade and Tavares, 2005. The classification of each instance (text) by each method (classifier) was compared to the expert annotation taking into account our 3 bands of text complexity (3rd year, 4th/5th years and 6th/7th years). For the dichotomous models (the first two and the latest models), the classification was considered as correct, if it was equal to the expert annotation, or as incorrect, otherwise. The difference for the polytomous model is at the incorrect classifications: they were ordered according to the distance between the method and expert classification bands. The fourth IRT model considered in this work was proposed to allow correlations among responses for the same learning algorithm. All analysis were performed in R software. Marginal maximum likelihood and expected a posteriori methods were adopted for item and ability parameter estimations, respectively.

3. Results

A partial picture of the results are presented for an illustration of the application of 2-PL model to the data. The interpretation is divided in two parts: (i) the instance difficulties and discrimination powers and (ii) the ability of the classifiers. The difficulty parameter estimates vary from -427 to 110, with 1st and 3rd quartiles equals to -3.6 and 0.5, respectively, and median standard error equals to 0.5. These values are well correlated to the percentage of classifiers that predict the instances correctly. The extremely high absolute values (greater than 4, for instance) correspond to classifiers that have a very low (or high) percentage of correct classification. The discrimination parameter estimates of an instance can be used to indicate if the instance is useful to distinguish between strong and weak classifiers for a problem. The obtained values varies from -2.6 to 5.4, with standard errors varying from 0.1 to 1.5 (95th percentile equals to 0.5). The negative values mean that these instances are most frequently well classified by the weakest classifier, identifying particular situations such as wrongly labeled instances, overfitting or noisy instances (the ones that are in regions of the instance space dominated by the other classes). The identification of these key instances is very useful in the context of supervised machine learning problems.

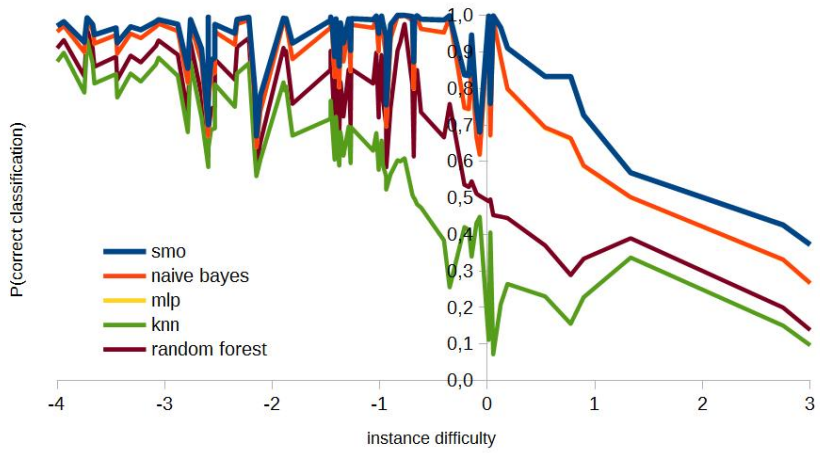

Figure 1: Frequency of correct responses of four classifiers with ability estimates equals to 2.6 (smo), 1.36 (naive bayes), 1.35 (mlp), -0.68 (knn), and 0 (random forest).

The estimation of the latent trait reflects the comparison of classifier qualities. Their values varied from -3.9 to 2.1, with a maximum standard error of 0.3. Figure 1 depicts the probability of correct classification for some classifiers in function of difficulty parameters. Instances with negative or low slope (less than 0.3) were

excluded for better clarity.
It is important to note that the number of classifiers might not be enough for obtaining good estimates, once the literature refers to 1,000 subjects as necessary to have an acceptable estimation. However, the principal aim of this situation is to estimate classifier abilities, and the large number of instances guarantee low standard errors for latent trait estimation, as corroborated by present results.
The other 3 models that will be applied to the data have interesting aspects to be concerned. In the 3-PL model, for instance, some effort should be made on the guessing parameter interpretation, which does not follow the intuitive idea from Psychometrics. It might be interpreted as an extra degree of freedom to fit the logistic models, but not linked to the number of classification categories (Martínez-Plumed et al., 2016). The application of graded response models to these data is interesting in the sense of consider classification not only as correct or incorrect but also taking into account the magnitude of classification error (distant from 1 or 2 bands from correct classification). Finally, the application of an IRT model for repeated measures in subjects (classifiers) will allow comparisons among algorithms families in general, more than algorithms with specific values attributed for respective parameters.

4. Conclusions

In this work we have investigated the use of IRT for the analysis of instances and classifiers of machine learning. We understood the meaning of the item parameters in reflecting instances with noise or overfitting. We were able to rank machine learning algorithms according to their abilities to correctly classify a text, depending on its difficulty and discrimination capacity. We also able to identify problematic instances that can be removed from dataset in order to better train less powerfull (but cheaper) machine learning algorithms in context of systems with limited processing and memory resources.
We proposed more sophisticated IRT models to consider dependency among classifications coming from the same algorithm family and polytomous model to differentiate classifications more or less apart from the real one.

References

aaaa, aaaa.

Aluísio and Gasperin, 2010 Aluísio, S. M. and Gasperin, C. (2010). Fostering digital inclusion and accessibility: the PorSimples project for simplification of Portuguese texts. In *Proceedings of the Young Investigators Workshop on Computational Approaches to Languages of the Americas (NAACL-HLT-2010)*, pages 46–53. Association for Computational Linguistics.

Andrade and Tavares, 2005 Andrade, D. F. and Tavares, H. R. (2005). Item response theory for longitudinal data: population parameter estimation. *Journal of Multivariate Analysis*, 95:1–22.

Birnbaum, 1968 Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Infering an Examinee's Ability. In *Statistical Theories of Mental Text Score*. Addison-Wesley, Reading,MA.

Feng et al., 2010 Feng, L., Jansche, M., Huenerfauth, M., and Elhadad, N. (2010). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010): Posters*, pages 276–284. Association for Computational Linguistics.

Graesser et al., 2011 Graesser, A. C., McNamara, D. S., and Kulikowich, J. M. (2011). Coh-metrix providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5):223–234.

Graesser et al., 2004 Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Hall et al., 2009 Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Lennon and Burdick, 2004 Lennon, C. and Burdick, H. (2004). The lexile framework as an approach for reading measurement and success. *Electronic publication on www.lexile.com*.

Martínez-Plumed et al., 2016 Martínez-Plumed, F., Prudêncio, R. B. C., Usó, A. M., and Hernández-Orallo, J. (2016). Making Sense of Item Response Theory in Machine Learning. In *Proceddings of the 22nd European Conference on Artificial Intelligence (ECAI-2016)*, pages 1140–1148.

Martins et al., 1996 Martins, T. B., Ghiraldelo, C. M., Nunes, M. d. G. V., and de Oliveira Junior, O. N. (1996). *Readability formulas applied to textbooks in brazilian portuguese*. Icmsc-Usp.

Samejima, 1969 Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*, 34(17):1–100.

Stenner, 1996 Stenner, A. J. (1996). Measuring Reading Comprehension with the Lexile Framework.

Witten and Frank, 2005 Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.