Determining Item Paramaters for Developing Better Testing Systems

Arbaz Khan arbazk@cse.iitk.ac.in
Prof. Harish Karnick hk@cse.iitk.ac.in

Department of Computer Science and Engineering
Indian Institute of Technology, Kanpur

CS497 report
November 14, 2012

Abstract

Modern day university entrance examinations witness large scale participations, all competing for comparatively very low number of seats. Such a tough scenario not only enforces a lot of pressure to the students but also sets challenges for the paper setters. The aim is to come up with a set of questions (items) which evaluated eventually would best serve the purpose of examination to extract qualitative information from the scores. In this work, we aim at predicting beforehand item parameters of toughness and discrimination using measures on cognitive faculties that a student uses to answer a question. We would further use the predictions to help reduce the flaws contained in modern day large scale offline testing systems.
Keywords

1 Introduction

Continuing our previous work of better student evaluation by predicting student's answer, we make an attempt over improving student evaluation by targeting present day testing systems. There are numerous large scale offline testing systems for admissions in undergraduate and graduate studies, most of which are objective based competitives. Hundreds of thousand students compete for a place among few top thousand students. Restricting the time available for testing, there can only be limited questions put forth to the mass. This leads to a situation of a very tough neck to neck competition amongst the students. IIT-JEE conducted by Indian Institutes of Technology is the most well known examination exposing this scenario. The results are such that even at the very top ranks (1000-2000), there are about a couple hundred of students at the same score. Situation gets even worse as we go to bottom ranks where one could drop or add thousands of ranks with a difference of about unit marks. As the purpose of the examination is to select the few best in the large mass of students, it becomes important for the paper to contain a set of questions which extracts the best students and also explores the discrimination between them minimzing the variation in the number of positions per unit difference of marks.

We tackle with the problem of better testing system by analysing the case of IIT-JEE. We first introduce the estimation methods that we use for item parameters of question toughness and question discrimination to anticipate the response to an item. We describe the power discrimination parameter holds in such examinations and how does it critically help in deciding the right set of items to be included in the test. We then introduce the challenges in predicting this techniques in testing systems and further explore the techniques to meet the challenges and predict these parameters. The discrimination parameter of an item defines how strong the item is in discriminating persons with different abilities. Although, literature provides multiple number of ways to estimate the discrimination parameter (in [2] and [1]), we infer a new estimation technique specific to the context of offline large scale testing techniques. We try to use these predictions to come up with a subset of questions from the original question set that could preserve the best in the crowd yet clearly discriminate them. Finally we conclude by presenting the results as to how the predictions perform as compared to actual values.

1.1 Related Work

Item Response Theory (IRT) provides comprehensively established literature with works of decades. Although the focus of attention has been mainly in estimation of item parameters. Both question toughness (item difficulty) and discrimination parameters have been mentioned in the Item Response Theory(IRT) with mutiple estimation techniques (Bock and Aitkin [1981], Baker and Kim [1981], Fan [1998] and Lord [1986]) in order to extract the true esti-

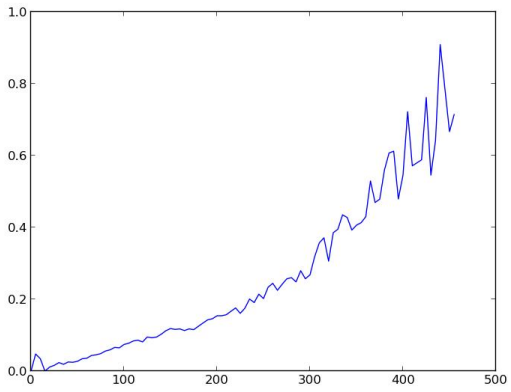


Figure 1: Item Characteristic curve for item with difficulty index = 0.068 and discrimination index = 0.2.

mate of a person's ability. A work has been done very recently (M Matteucci and Veldkamp [2012]) in order to predict item parameters by utilizing the empirical information from items. Also, item covariates have been tried to infer from surveys but one cant naively apply these techniques in predicting item parameters for large scale testing systems.

2 Estimating item Parameters

2.1 Item Difficulty

Item difficulty has been estimated using various techniques in IRT to extract the true ability of a person rather than relying upon the observed score which classically is number corrects upon total. However, it has been found that IRT estimation techniques highly correlate with classical scoring. Also, in this context when we use item difficulty as a parameter to aid paper setting and item selection, it helps if one could easily visualize the impact of this parameter. Hence we estimate item difficulty as :

$$\textit{Item Difficulty Index} = \frac{\textit{Number of correct responses}}{\textit{Total responses}}$$

2.2 Item Discrimination

Item discrimination as mentioned earlier is the strength of an item in discriminating persons with different abilities. Elaborating in this context, it is the measure of how well does the item serve to discriminate students of higher abilities with the lower ones. Since, we intend to select very few number of students as compared to the total participations, the discrimination index of an item should give the extent of success of the item in identifying the top ranked students. In most of the examinations, the evaluators decide certain cutoff total marks to adjudge selection to the students. Here,

QuestionID	Difficulty Index	P_C - P_{-C}
111	0.0374	0.073
212	0.04	0.105
101	0.259	0.353

Table 1: Correlation: Higher the toughness lower the discrimination index.

we use this cutoff marks as the measure to distinguish the students of higher ability with those of lower ability. The greater the difference in probabilities of giving correct response to the item by the above cutoff students as compared to the lower cutoff students, the higher the discrimination index. Hence, comes the estimation of discrimination index as :

$$\textit{Discrimination Index} = P_C \text{ - } P_{-C}$$

where:
P_C is the percent correct responses to the item by students above cutoff marks
P_{-C} is the percent correct responses to the item by the rest
Discrimination index holds a lot more utility as compared to difficulty index in selecting the right questions for an examination. We elaborate in the following section how can one exploit pre-knowledge about an item's discrimination.

3 Determining Item Parameters

3.1 Problem Statement

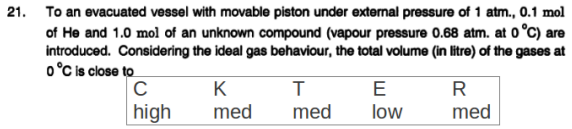
Given a set of items to be fed in a testing system, we aim to predict item parameter indices of difficulty and discrimination and use the predictions to extract the subset of items to help identify the students with higher abilities as early as possible subsequently creating a scope of further discrimination amongst them.

3.2 Natural Dataset

The natural dataset for a competitive exam contains the following information:

- Student ID
- Question ID
- Question Type (e.g. single correct, match the columns)
- Question Track (e.g. Physics, Mathematics)
- Student Response
- Correct Response
- Maximum marks

The number of items(questions) available in the data set were 147, out of which 140 were legit and the rest ended up being awarded correct to all participants. The number of students being evaluated were 407752.


21. To an evacuated vessel with movable piston under external pressure of 1 atm., 0.1 mol of He and 1.0 mol of an unknown compound (vapour pressure 0.68 atm. at 0 °C) are introduced. Considering the ideal gas behaviour, the total volume (in litre) of the gases at 0 °C is close to

C	K	T	E	R
high	med	med	low	med

Figure 2: Assignment of cognitive attributes to the values by an expert.

3.3 Challenges

As mentioned before, the previous techniques of item parameter prediction rely upon either empirical information contained in the item or treat these parameters as random. Also, short scale surveys are taken to get an estimate of the parameters beforehand. In our case, in an examination like IIT-JEE one cant get a survey done for the items as this breaks the strict confidentiality maintained only between the paper setters. And neither can one rely upon empirical information contained in the items as these exams are made to have a decent enough variations every year preserving their quality. Howver, one could aid the predictions by use of experts but naively discrimination indices of items cant be predicted by experts Duyugu. This easily intuitive as there underlies a lot of complexity in meeting the conflicts of opinions and also on the degree of subjectivity put on the item. What is a seemingly tough question could be judged moderate by one and very difficult by other. Hence, we need a technique to utilise the information provided by experts in item parameter prediction.

3.4 Approach

Since item parameters are somehow related to the measure of the student abilities, what if we could augment abilities with each item and let the experts predict the extent of each of these required to answer the question. We start by defining the item in terms of the esimates of the abilities. Hence, measures of abilities of each of these becomes our feature set. We take our estimates of the extent of these abilities from experts, responses of which can only be either *low*, *medium* or *high*. The extents of each of these abilities becomes our new data instance. To make a prediction now on the parameters, say item difficulty, we use regressiontree model attempting to make a regression on the item difficulty. So build regression trees for each of the experts and train the regressor by trainining the experts on known item sets. Consequently each regressor outputs us the prediction for item difficulty. To combine the results, we use another regressor which takes as input the predictions of each of the regression trees and in turns outputs the final prediction. The approach is pictured in Figure 3 and each aspect is broadly covered in the subsequent subsections.

3.4.1 Describing the Feature Set

The feature set of our new dataset would be the cognitive attributes involved in answering a question. For the purpose of experimentation, we started with 5 cognitive attributes defining a question. The five attributes we used span the cognitive skills of understanding concepts(**C**), applying knowledge from memory(**K**), processing of mind involved in making calculations (**T**), common sense of eliminating choices (**E**) and logical reasoning (**R**). The experts measure on each of these is one of *low*, *medium* and *high*. These attributes may or may not comprehensively cover the complete set of faculties involved but these certainly form the most significant attributes of answering a question. The number of dimensions, how many to actually set up would require experimentation results to support them as not always do regression techniques preform best with higher dimensional data. Sample data instance from one of our experts on one of the items from the set can be seen in Figure 2.

3.4.2 Introducing the regression layers

We used to layers to efficiently predict the item parameters. Layer1 (Figure 3) consists of regressors applied over each of the expert's data. The above mentioned features would be the features of the regression tree. Class label would be the item parameter's true value. The regressor at Layer2 combines the predictions provided by different experts so that the problem of solving conflicts of subjectivity of thought to the regressor to reflect. So when a strict expert adjudges moderate toughness everytime to a very tough question, a good regressor (at Layer -2) can capture this strict behaviour of one of its regressor and take it into account while making future predictions. Hence the feature set of Layer-2 regressor is the outputs from each of the regressors in Layer1 and class label is parameter prediction.

3.4.3 Training the system

Since, our dataset had a small sample size of 140 items, training the regressors becomes an issue especially when we have 2 layers of regression as mentioned above. The training set is not divided any further and all 120 items are fed to Layer-1. While training, each of the regressors in Layer1 is cross validated with leave one out (Loo) cross validation technique. The prediction outputs for each validation and for each regressor is stored along with the true prediction and this is used to train the regressor at Layer-2. The use of **Loo** cross validation technique provides necessary enough training data for Layer-2 or else a test-training split further at Layer-2 with some 20 odd training instances for Layer-2 would not be able to avoid fitting errors.

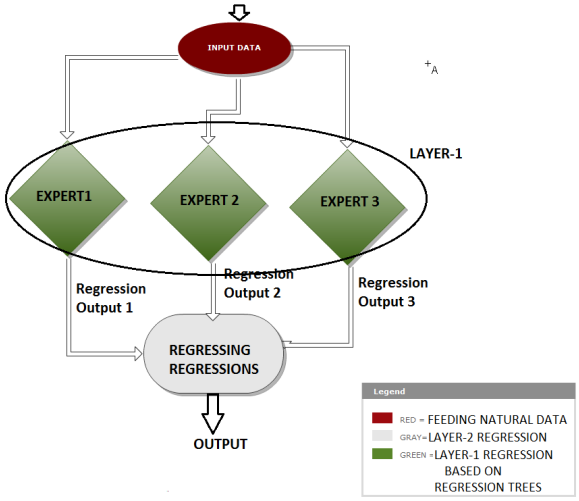


Figure 3: Explaining how Prediction System Works.

Passive zone for
High Ability Students

232	1	0	0	0.065
145	0	0	0	0.057
134	0	0	0	0.044
105	0	0	0	0.022
166	1	1	1	0.006
2194	0	0	0	0
253	0	0	0	0
152	0	0	0	0
155	1	1	1	-0.01
252	0	0	0	-0.02
216	0	0	0	-0.02

149	1	1	1	0.59
245	1	1	1	0.588
115	0	1	1	0.586
124	1	0	1	0.58
249	1	1	0	0.574
224	1	1	1	0.574
127	1	1	1	0.572
154	1	1	1	0.571
229	1	1	1	0.567

Active Zone

Figure 4: An analysis of performance of three students at 250 marks helps visualize the correlation between High Discrimnination Index and success of High ability students. The last column is discrimination indices of the items. Passive zone is where students fail often.

4 Utilizing the Predictions

In this section, we deal with how can one utilize the predictions of difficulty index and discrimination index to set up a good question paper. From the given item set, we need to choose the smallest subset of questions which could within themselves identify the best students. This requires an insight into the power of discrimination index in identifying higher ability students.

4.1 The Potential of Discrimination Index

We start with an analysis of dataset in hand. We had computed discrimination indices for all the items in the dataset and sorted each in the descending order of their discrimination indices. The item with higher discrimination index is set to contain more the students with higher ability as we defined it so. So the intuition is that optimal subset we are looking for must be all the items with top K discrimination indices. To establish our intuition, we used various sizes of item sets each time choosing top K discriminative items and evaluated the students only on this subset of items and ranked accordingly for upto a certain number, N [1] of ranks. Results are summarized in Table 2. Threshrank is the threshhold for discrimination index while choosing the subset of questions. Containment is the percent of students contained in the new rank estimates which were originally present in the top N positions. The results hint that with about only 50 items (35.7% of itemset), one could contain the original 85% students. The rest of the items practically hold less significance in identifying the best candidates. These items were those which have low discrimination index and only distribute the marks to the medium ability students. These are the items which are either very easy so that students with all ability succeed on it, or are very tough which not many succeed. The lower ability students get to score better on these items than the moderate ability students probably due to guessing (as objective itemset) or due to lack of knowledge about the complexitites of such a question. One can visualize such items through the Item Characteristic Curves in Figure-5.

4.2 Choosing the best set of items

Having pointed out the flaws in the current large scale offline testing systems, the question is to how can one decide upon choosing the best set of items in pre-testing situation. For that, one can choose such a prediction system to the aid. One can feed in more than required number of items in the prediction system. The system if reliably enough can predict the top discriminative items, then this is the set of items one can choose from. The choice of containment has tradeoff with the size of itemset selected. Consequently, item difficulty is the parameter which creates a discrimination amongst the higher ability students. The tough questions do have a very low discrimination index overall but within the high ability students, it acts as the discrimination parameter. Item difficulty prediction has a great correlation with discrimination index and also can be used to validate a prediction manually. If the itemset contains larger number of high discrimination index items, it potentially has the ability to better distinguish the abilities finely.

[1] This number is predecided by the examiners according to the number of available seats.

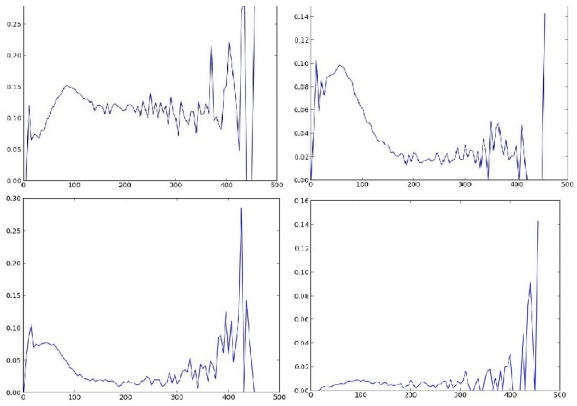


Figure 5: Item Characteristic Curves Of Items with High Difficulty index and very low discrimination index.

Threshrank (tr)	Containment(%)
20	72.22
25	76.62
40	83.53
50	85.73
70	90.17
100	94.2
125	96.60

Table 2: Containing the top ranked students using threshrank number of questions.

5 Experimental Results

For the purpose of prediction, item difficulty was discretized into 10 difficulty levels(0-9, 0-toughest) and item discrimination was discretized into 5 discrimination levels (0-4, 4-most discriminative). Each of these divisions were equal in terms of the number of items contained. The prediction training test system was then run on these estimates.

We used the following techniques for regression:

Layer:1

M Matteucci and Veldkamp [2012] Supports the use of regression trees for predicting item parameters and it has been found to be in correlation with experimental results over different classifiers. The optimal parameters for regression tree were SSE based error calculation with binary node each labelled with a question of type $K < 4$?. Termination parameters wre SSE Threshhold set to 0 and minimum length of a leaf node to be 3.

Layer:2

Techniques of Bagging, K-star and SVM based regression were used from Weka toolkit. (Hall et al. [2009]). The parameters used were mostly default parameters, which if devaiated are stated in the results tables 3 and 4.

Technique at Layer-2	Mean Absolute Error	Correlation Coefficient
Bagging	2.19	0.366
SVM	2.11	0.3964
K-Star	2.17	0.384

Table 3: Results on cross validation for Item Difficulty prediction. Layer-1 regressor is regression tree in all cases

Technique at Layer-2	Mean Absolute Error	Correlation Coefficient
Bagging	1.12	0.37
SVM(c=2)	1.25	0.09
K-Star	1.15	0.329

Table 4: Results on cross validation for Item Discrimination prediction. Layer-1 regressor is regression tree in all cases

Apart from using cross validations, for once to visualize the power of the prediction system (Table 5), we split the total dataset in 3 portions. 120 data sets were used to train Layer-1 which were tested over 12 data sets. These 12 outputs formed the training set of Layer-2 and it was tested on the 8 data sets to create a 66-33% training-test split. The results were very much accurate with 11% mean absolute error using K-Star at Layer-2. Although this can be misleading but motivates to further improvise the predictions.

6 Conclusions

The results speak that with the existin system we can predict item parameters with an average error of two levels of difficulty and one level of discrimination. This is motivating as an average of one level of discrimination implies correct prediction of round about upto 28 positions the discriminative rank of an item in top discriminative items. This largely speaks of the utility of the system in choosing the most discriminative postions.

But still one could think of a lot more improvement in the prediction rate as the pattern indicating a weak expert was obviously present which was not detected. We would focus upon improving this prediction performance so that the system could be put to some worth in modern day testing sytems.

References

F Baker and S Kim. Item response theory parameter estimation techniques. *Psychometrika*, 46:157–162, 1981.

R Bock and M Aitkin. Marginal maximum likelihood es-

Instance #	Actual Difficulty level	Predicted Difficulty level
1	2	1
2	6	8
3	5	4
4	9	6
5	0	1
6	1	1
7	8	8
8	7	8

Table 5: System result using training-test split evaluation

timation of item parameters: Application of an em algorithm. *Psychometrika*, 46:157–162, 1981.

Anil Duyugu. The prediction of item parameters based on classical test theory and latent trait theory.

Xitao Fan. Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Psychometrika*, 3:157–162, 1998.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009. ISSN 1931-0145. doi: 10.1145/1656274.1656278. URL `http://doi.acm.org/10.1145/1656274.1656278`.

F.M. Lord. Maximum likelihood and bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, 23(2):157–162, 1986.

S Mignani M Matteucci and P Veldkamp. The use of predicted values for item parameters in item response theory models: an application in intelligence tests. *Journal of Educational Measurement*, 2012.