Auto-Tagging for Massive Online Selection Tests: Machine Learning to the Rescue

S.Krithivasan, S.Gupta, S.Shandilya, K.Arya, K.Lala
Department of Computer Science & Engineering
Indian Institute of Technology Bombay, Powai
Mumbai - 400076, India

Abstract*—Difficulty Level of a question is relative to that of other questions in a test and also to the test takers; hence manually assigning Difficulty Level tags may not be accurate. There is a need to infer them from historical data pertaining to the performance of students in a test. e-Yantra Robotics Competition (eYRC) is an annual competition having around 5000 teams (20,000 students) registering in the latest edition of the competition, eYRC-2015. All four team members take a test simultaneously and each individual gets questions which are different but have a similar Difficulty Level. A Question Bank containing 1800 unique questions from 3 subjects - Aptitude, Electronics, and C-Programming – is used to generate question sets each having 30 questions. It is a challenge to ensure that each set contains questions of similar Difficulty Levels tagged manually as Easy, Medium or Hard. In this paper, we discuss a learning algorithm called Weighted Clustering that can automatically tag questions by analyzing the performance of students. We used this algorithm to analyze the performance data in eYRC-2014 for 614 questions from the Question Bank; we found that Manual Tagging accuracy was 44%. We re-tagged questions with Suggested Tags resulting from our analysis and used them again in eYRC-2015. When we applied the algorithm to the performance data in eYRC-2015, we found that the accuracy of tagging had significantly improved to 67%.

***Index Terms*—Machine Learning, Online Testing Environment, Robotics Competition, Selection Test, Weighted Clustering, e-Yantra.**

I. Introduction

A computer-based online test is different from a traditional testing environment in many ways. As discussed in [1], some of the advantages include reduced cost, accurate scoring and scalability. This however raises various other challenges as discussed in [2]. One of the major challenges is to ensure that the test is fair in the sense that all students taking the test are presented with questions of a similar Difficulty Level. Ensuring fairness becomes especially necessary if a uniform marking scheme is in use.

In this paper we present the methodology used in a computer based online selection process for a nation-wide robotics competition [3][4]. A total of 12,428 students (3107 teams, each having 4 members) registered for the e-Yantra Robotics Competition in 2014 (eYRC-2014), and 19,568 students (4892 teams) registered in eYRC-2015. Registered teams appear in an online Selection Test. All four team members take the test simultaneously and each test consists of questions which are different but have similar Difficulty Level. Each team member answers 30 multiple choice based questions, referred to as a **set**, drawn randomly from a Question Bank. Teams are shortlisted based on average of marks scored by all four members.

In this paper we present answers to the following questions:

1) What are the properties of sets and how do we create sets such that they have these properties?
2) How do we tag questions in the Question Bank with a Difficulty Level?
3) What accuracy level can we achieve by assigning Difficulty Level tags to questions manually?
4) How can we improve tagging by applying an auto-tagging algorithm?

II. Selection Test: Process Flow

A. Creating the Question Bank: Manual Tagging Process

To start with, fifteen Question Setters independently created a certain number of questions each and tagged these questions with a **Difficulty Level** tag. Typically, questions that require:

1) direct application of concepts and formulae are categorized as **Easy**.
2) extra processing such as application of assumptions and/or prior knowledge are categorized as **Medium**.
3) derivations and application of logic that involve multiple steps to arrive at a solution are categorized as **Hard**.

Corresponding to every Question Setter is a Question Reviewer, who checks each question for correctness and validates the Difficulty Level tag assigned by the Question Setter. Any conflict regarding the tags is resolved via discussion. The tags thus assigned are called **Manual Tags**.

B. Mapping Questions to Sets

In eYRC-2015, 1800 questions are mapped into 120 sets where each set contains 30 questions - 10 questions from each of the three categories - Aptitude, Electronics and C-Programming. Within each category of a set the questions are mapped such that there are 3 Easy questions, 4 Medium questions and 3 Hard questions. This is illustrated in Figure 1. Sixty sets were created by taking each of the 1800 questions exactly once. These 60 sets are called **Unique Sets**. The remaining 60 Sets, called **Extended Sets**, are derived from the Unique Sets.

Assuming that Difficulty Level tags assigned manually are correct, this procedure ensures that all the sets are balanced i.e. they have similar overall Difficulty Level.

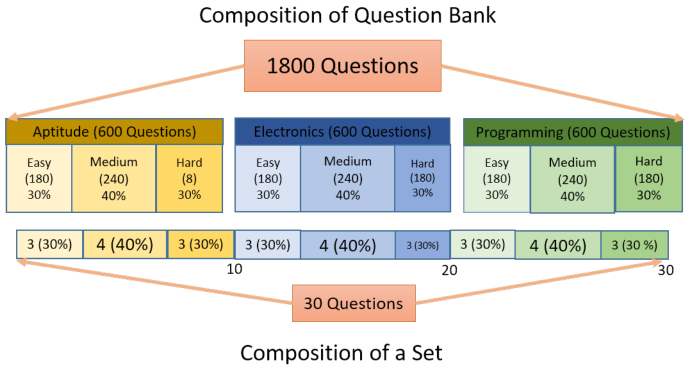


Figure 1: Composition of Question Bank and Set

However, since the tags are given independently, a question, which is Hard relative to the questions created by a Question Setter might be an Easy question when all 1800 questions in the Question Bank are considered together (relative nature of Difficulty Level of questions). Further, the manual tag assignment is a subjective process and hence it is not very accurate.

eYRC Selection Test is designed in such a way that students from different engineering backgrounds find it equally challenging. A number of variables discussed in Item Response Theory (IRT)[5] were eliminated using statistical hypothesis testing to formulate a simpler model with less noise that works well in our case. We tackled the problem from a Machine Learning perspective. The task of assigning Difficulty Level tags to questions is essentially an Unsupervised Learning [6] problem because we do not have a set of questions with known true Difficulty Level tags to guide Supervised Learning [7] algorithms like Logistic Regression [8], SVM [9] etc. Hence to validate the accuracy of the tags we need to perform Unsupervised Clustering of data.

III. Validating Manual Tagging: Weighted Clustering Algorithm

Weighted Clustering algorithm is a derivative of the well known k-Means Clustering [10][11] algorithm. Both the algorithms partition the data (questions) into clusters (Easy, Medium, Hard). A cluster is a group of similar data items based on some **similarity metric** [12](for example: Euclidean distance between two data points is a similarity metric). To use the similarity metric, the data must be represented in a form on which the chosen similarity metric (Euclidean distance) can operate. This representation of data is called **Feature Representation of data**.

We define the following terms to explain the Weighted Clustering algorithm:

1) **Data Point:** A Data Point refers to a 3-tuple (Student ID, Question ID, Marks). For example, if a student with Student ID 90, answers a question with Question ID 564 incorrectly then the tuple generated is (90, 564, -1). Note that Marks can take only three distinct integer values: +3 (correct answer), 0 (not attempted), -1(incorrect answer). Since 30 questions are served to each student, every student will generate 30 Data Points.
2) **Source:** In general, Source refers to any entity that can generate Data Point(s). In the context of the Selection Test, Source refers to a student.
3) **Weight:** Different Sources that generate Data Points have different reliability. We assume that a high scoring student is more reliable than a low scoring student, in the sense that s/he is less likely to make random guesses. Weight is a real number in the range (0, 2) assigned to each Source, based on Mean Normalized Marks obtained by the students in the test. A high scorer gets a higher Weight (close to 2) while a low scorer gets a lower Weight (close to 0).
4) **Feature:** Feature refers to a quantitative measure based on which a learning algorithm makes its decision. It is a numeric value for each question, derived from the performance data. For example, the fraction of students who have solved the question correctly may be a Feature. A list of Feature values about a question form the Feature Representation for that question.

A. Weighted Features

Weighted Clustering algorithm uses two Features for every question:

1) $F_k^{(1)}$: Weight adjusted fraction of students who answered the k^{th} question correctly. If $S = \{j | j^{th}\, student\, was\, served\, the\, k^{th}\, question\}$, then $F_k^{(1)}$ can be calculated using Equation 1. Here $f_k^{(1)}(.)$ is a function that takes as input a Data Point $D_k^{(s)}$ and evaluates to 1 if s^{th} student answered the k^{th} question correctly and to 0 otherwise. w_s refers to the Weight assigned to the s^{th} student.

$$F_k^{(1)} = \frac{\sum_{s \in S} w_s f_k^{(1)}(D_k^{(s)})}{\sum_{s \in S} w_s} \tag{1}$$

2) $F_k^{(2)}$: Weighted average of marks of students who did not attempt or were not able to solve the k^{th} question correctly. Defining S and w_s same as in $F_k^{(1)}$, $F_k^{(2)}$ can be calculated using Equation 2. Here m_s refers to the total marks obtained by s^{th} student and $f_k^{(2)}(.)$ is the function that takes as input a Data Point $D_k^{(s)}$ and evaluates to 1 if s^{th} student did not attempt or was not able to solve the k^{th} question correctly and to 0 otherwise.

$$F_k^{(2)} = \frac{\sum_{s \in S} w_s m_s f_k^{(2)}(D_k^{(s)})}{\sum_{s \in S} w_s} \tag{2}$$

Features $F_k^{(1)}$ and $F_k^{(2)}$ are called Weighted Features. They are calculated for every question k in the Question Bank. Intuitively $F_k^{(1)}$ will be high for Easy questions and low for Hard questions because more test takers will get an Easy question correct. Similarly $F_k^{(2)}$ will be low for an Easy question and high for a Hard question because the average marks of test takers who were unable to answer an Easy question will be low.

The Weighted Clustering algorithm iteratively finds better clusters with an objective of decreasing variance within clusters. Here variance refers to the mathematical concept of variance based upon the Feature Representation of data and similarity metric. Intuitively the algorithm finds clusters such that “similar” data belongs to the same cluster taking into account the Weight assigned to each distinct Source.

B. *Assignment of Semantic Labels to Clusters*

Clustering algorithm groups similar data into clusters. We need an ordering of these clusters as Easy, Medium or Hard. To assign semantic label to clusters we calculate first quartile, median, third quartile and mean value for each **cluster-Feature pair**. Manual assessment of this data is done to assign a semantic label to each cluster based on intuitive interpretation of Features.

IV. Validation of Difficulty Levels

For eYRC-14 Selection Test we had manually assigned tags for all the questions in the Question Bank. We refer to these tags as Manual Tags (MT). We applied the Weighted Clustering algorithm to the marks scored by the students in the eYRC-14 Selection Test. This provided us with Suggested Tags for the questions used in eYRC-2014. We refer to these as ST_{2014}. In order to ensure integrity of the selection process, in eYRC-15, some questions in the Question Bank were deleted, modified or added. After all changes were made, 614 questions retained their original semantics as in eYRC-14 Question Bank. We applied the Weighted Clustering algorithm to the marks scored by students in eYRC-15 Selection Test to obtain Suggested Tags for these 614 questions. We refer to these as ST_{2015}. For all our subsequent analyses we will use the tags of these 614 questions only.

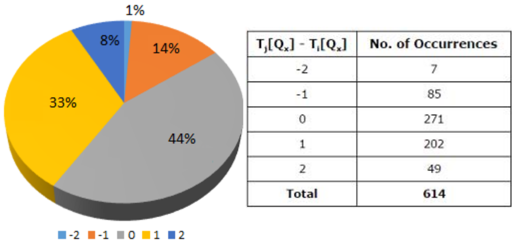
We use the following notation in all the analyses presented for comparison of tagged values obtained using two different **Tagging Methods i** and **j**, on a given set of questions:

- T_i and T_j represent tagged values of a given set of questions using Tagging Method i and Tagging Method j respectively.
- A question Q_X can be tagged as Hard (having a value 2), Medium (having a value 1) or Easy (having a value 0). Suppose the tagged value of Q_X using i is Hard, denoted as $T_i[Q_X] = 2$, it is possible that it is tagged as either Hard, Medium or Easy by j. To compare i with j, we find the difference in Difficulty Levels ($T_j[Q_X] - T_i[Q_X]$). The different possibilities are illustrated in Table I.

Table I: Difference in Difficulty Levels Between Tagging Methods i and j

$T_i[Q_x]$	$T_j[Q_x]$	$T_j[Q_x] - T_i[Q_x]$
Hard (2)	Hard (2)	0
	Medium (1)	-1
	Easy (0)	-2
Medium (1)	Hard (2)	1
	Medium (1)	0
	Easy (0)	-1
Easy (0)	Hard (2)	-2
	Medium (1)	-1
	Easy (0)	0

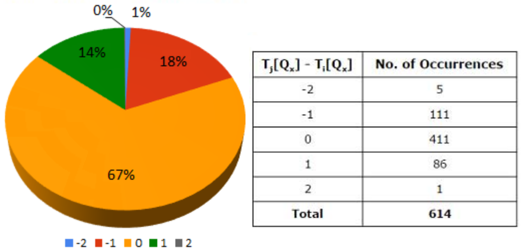
In the following analysis, $T_i = MT$ and $T_j = ST_{2014}$. The table inside Figure 2 lists the number of occurrences of the different $T_j[Q_X] - T_i[Q_X]$ values and for easy of understanding, these values are represented in pie chart.



$T_j[Q_x]$ - $T_i[Q_x]$	No. of Occurrences
-2	7
-1	85
0	271
1	202
2	49
Total	614

Figure 2: Percentage Occurrences of Difference in Difficulty Levels (T_i=MT and T_j=ST_{2014})

With reference to Figure 2, for 271 questions (44.1%) the tags given by the two Tagging Methods i and j matched. A total of 287 tags (85+202) are 1 level away, i.e. a difference of ±1. This constitutes 46.7%. 56 questions (49+7) are 2 levels away - note that this case can occur in two cases: (i) a Easy question is tagged as Hard or (ii) a Hard question is tagged as Easy. Both these cases are undesirable and equivalent to tagging a question **incorrectly**. In order to be fair to students taking the Selection Test, such that all sets are of similar Difficulty Levels, our goal is to increase the match (occurrences of 0) while decreasing number of questions tagged incorrectly.



$T_j[Q_x]$ - $T_i[Q_x]$	No. of Occurrences
-2	5
-1	111
0	411
1	86
2	1
Total	614

Figure 3: Percentage Occurrences of Difference in Difficulty Levels (T_i=ST_{2014} and T_j=ST_{2015})

To verify the goodness of the Weighted Clustering algorithm we compared ST_{2014} with the ST_{2015}. This analysis is presented in Figure 3. Note that T_i = ST_{2014} and T_j = ST_{2015}. For 411 questions (66.9%) the tags given by the two Tagging Methods i and j matched. A total of 197 tags (111+86) are 1 level away, i.e. a difference of ±1. This constitutes 32.1%. Only 6 questions (1%) are 2 levels away – tagged incorrectly.

V. Factors Affecting Accuracy

1) **Shifting:** Addition, deletion or modification of the semantics of questions in the Question Bank may affect the Difficulty Levels of the questions. For example, a question Q, that was originally Easy, becomes a Medium Difficulty Level question because of addition of other questions that were easier than Q. We term this as **Shifting** of tags.

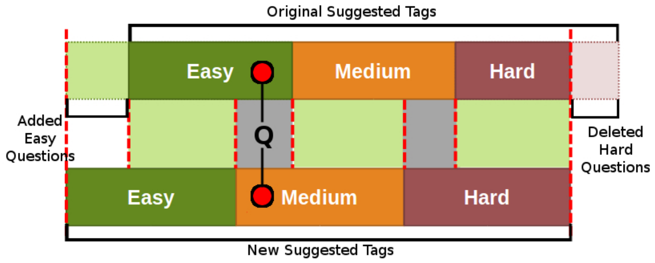


Figure 4: Shifting of Difficulty Level Tags

In Figure 4, the light green regions show overlapping tags (where $ST_{2014} = ST_{2015}$) while the gray regions show non-overlapping tags (where ST_{2014} and ST_{2015} differ). The gray regions are responsible for loss in accuracy.

2) **Switching:** Difficulty Level forms a continuous spectrum which we discretize into three distinct Difficulty Level tags. Due to this, questions present at the boundary between two levels can easily exchange their tags even due to slight variation in data obtained. We term this as **Switching** of tags. Figure 5 shows the number of questions that Switched their Difficulty Level tag by a single level in our analysis.

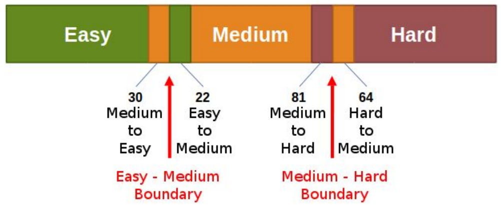


Figure 5: Switching of Difficulty Level Tags

Thus Shifting and Switching of tags contribute to the loss in accuracy of the Suggested Tags. In our case since 1186 questions out of the 1800 questions in the Question Bank were modified, Shifting and Switching are highly probable.

VI. Conclusion

The online testing environment offers more advantages over traditional testing environments. But without fairness, no other advantages matter. This paper proposes a solution towards a fairer online testing environment. We presented a Weighted Clustering algorithm to assign Difficulty Level tags to the questions. We estimated the accuracy of Manual Tags by comparing them with automatically Suggested Tags and the accuracy was found to be 44%. We also assessed the accuracy of Suggested Tags by running the same algorithm on the data obtained from eYRC Selection Test for the successive year. The accuracy was found to be 67%. Using the discussion on Shifting and Switching of tags and their impact on accuracy, we justified why the accuracy did not improve by a higher percentage.

Acknowledgment

Our thanks to Prof. Krithi Ramamritham for his feedback and Amiraj Dhawan, Sachin Patil and the e-Yantra team for their support in designing and conducting the Selection Test. Acknowledgment is also due to NMEICT, MHRD, Government of India, for funding the project.

References

[1] R. E. Bennett, “Online assessment and the comparability of score meaning,” *ETS Research Memorandum*, November 2003.

[2] M. I. Mohammad A Sarrayrih, “Challenges of online exam, performances and problems for online university exam,” *IJCSI International Journal of Computer Science Issues*, vol. 10, no. 1, January 2013.

[3] S. Krithivasan, S. Shandilya, K. Arya, K. Lala, P. Manavar, S. Patil, and S. Jain, “Learning by competing and competing by learning - experience from the e-yantra robotics competition,” *In IEEE Frontier In Education (FIE), 2014*, October 2014.

[4] S. Krithivasan, S. Shandilya, K. Lala, and K. Arya, “Massive project based learning through a competition - impact of and insights from the e-yantra robotics competition (eyrc - 2013),” *Technology for Education (T4E), 2014 IEEE Sixth International Conference*, December 2014.

[5] S. B. Kotsiantis, “Stata item response theory, reference manual, version14,” 2015. [Online]. Available: http://www.stata.com/manuals14/irt.pdf

[6] Z. Ghahramani, “Unsupervised learning,” *Gatsby Computational Neuroscience Unit, University College London, UK*, September 2004. [Online]. Available: http://mlg.eng.cam.ac.uk/zoubin/papers/ul.pdf

[7] S. B. Kotsiantis, “Supervised machine learning: A review of classification techniques,” *Informatica 31*, July 2007. [Online]. Available: https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf

[8] C. Shalizi, “Logistic regression,” *Department of Statistics, Carnegie Mellon University*. [Online]. Available: http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf

[9] C.-J. L. Chih-Wei Hsu, Chih-Chung Chang, “A practical guide to support vector classification,” *Department of CSE, National Taiwan University, Taipei 106, Taiwan*, May 2016. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

[10] “A tutorial on clustering algorithms.” [Online]. Available: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

[11] J. Macqueen, “Some methods for classification and analysis of multivariate observations,” *Fifth Berkeley Symposium*, vol. 67, no. 10, April 2013.

[12] A. R. Archana Singh, Avantika Yadav, “K-means with three different distance metrics,” *International Journal of Computer Applications*, vol. 67, no. 10, April 2013.