Int J Adv Eng Sci Appl Math (2018) 10(1):105–113
https://doi.org/10.1007/s12572-017-0202-9

IIT, Madras

Selection of automatic short answer grading techniques using contextual bandits for different evaluation measures

Shourya Roy[1] · **Arun Rajkumar**[2] · **Y. Narahari**[3]

Published online: 20 February 2018

Abstract Automatic short answer grading (ASAG) systems are designed to automatically assess *short* answers in natural language having a length of a few words to a few sentences. Many ASAG techniques have been proposed in the literature. In this paper, we critically analyse the role of evaluation measures used for assessing the quality of ASAG techniques. In real-world settings, multiple factors such as, difficulty level, and diversity of student answers, vary significantly across questions, leading to different ASAG techniques emerging as superior for different evaluation measures. Building upon this observation, we propose to automatic *learning* of a mapping from questions to ASAG techniques using minimal human (expert/crowd) feedback. We do this by formulating the learning task as a *contextual bandits* problem and providing a rigorous regret minimization algorithm that handles key practical considerations, such as, noisy experts and similarity between questions. Our approach offers the flexibility to include new ASAG systems on the fly and does not require the human expert to have knowledge of the working details of the system while providing feedback. With extensive simulations on a standard dataset, we demonstrate that our approach yields outcomes that are remarkably consistent with human evaluations.

Keywords Automatic short answer grading · Contextual bandits · Evaluation measures

1 Introduction

The task of automatically grading short answers, which are a few words to a few sentences long (everything in between fill-in-the-gap and essay type answers [1]), is referred to as *Automatic Short Answer Grading (ASAG)*. An example is shown in Table 1. Given a bunch of student answers to a question, machine learning and natural language processing (NLP) based ASAG techniques predict scores for each student answer with respect to the instructor provided model answer. Goodness of a technique is measured based on how closely the predicted scores for students match with the groundtruth scores. Towards that in ASAG literature, a range of evaluation measures have been used, albeit without much analysis of their suitability across different types of questions. A variety of values of absolute error measures such as mean absolute error (MAE) and root mean square error (RMSE); various correlation coefficients such as the ones due to Pearson and Spearman; as well as confusion matrix based measures such as Cohen's κ and F-measure across prior literature have been employed. The multitude of such evaluation measures has made effective comparison of ASAG techniques difficult, as emphasized by the recent survey papers by Roy et al. [2] and Burrows et al. [1].

While appropriateness of different evaluation measures for various machine learning and NLP tasks has been studied in several prior research papers, ASAG has certain

✉ Shourya Roy
shourya.roy@aexp.com

Arun Rajkumar
arun.rajkumar@conduent.com

Y. Narahari
hari@csa.iisc.ernet.in

[1] Big Data Labs, American Express, Bangalore, India

[2] Conduent Labs, India, Bangalore, India

[3] Indian Institute of Science, Bangalore, India

Table 1 Example of question, model answer, and student answers from an undergraduate computer science course [3]

Q-2	What stages in the software life cycle are influenced by the testing stage? (5)
Model Ans	The testing stage can influence both the coding stage (phase 5) and the solution refinement stage (phase 7)
Stud#1	Directly: refining, coding. Because Refining is right before the Testing Phase and Coding is right after the Testing Phase. Indirectly: Production, Maintenance. Because Refining occurs before these last two stages in the Software Life Cycle. (5/5)
Stud#2	Refining the solution, Production and Maintenance are all influenced by the Testing stage. (3/5)
Q-4	Where do C++ programs begin to execute? (5)
Model Ans	At the main function
Stud#1	At the main function int main() (5/5)
Stud#2	Main (5/5)

atypical characteristics warranting attention. An ASAG task is composed of a number of, typically independent, questions. The questions vary with respect to factors such as difficulty level, maximum attainable score, diversity of students' answers, nature of model answers etc. These factors affect performances of different ASAG techniques and consequently, the best among a competing group of techniques is not expected to be the same for all questions. In the next section, we show an example how unsupervised ASAG techniques (which rely on effectively capturing model and a student answer's similarity as a measure of the later's score) performance vary across questions. Moreover, the best technique for a question depends on the evaluation measure used for assessing their performances.

Motivated by the above and as the second contribution of this paper, we propose to intelligently choose ASAG techniques independently for different questions for overall best performance. Our endeavor is to learn a mapping from a set of questions to a set of ASAG techniques based on minimal feedback from human experts. Intuitively, the mapping would depend on characterizing factors of questions (difficulty level, nature of answers etc.) and will vary with the underlying evaluation measure. We model the learning task as an elegant contextual bandits problem where feedback for the initial questions is obtained using an expert with a specific evaluation measure. Given a question and a bunch of candidate techniques, the expert/crowd predicts scores and communicates the best among them. Over time the contextual bandits algorithm is expected to learn the expert's preference and would be able to predict the best technique given a question. We exploit a set of generalizable unsupervised (not requiring groundtruth scores) features extracted from model and student answers, to represent each question. In terms of the standard terminology, these extracted features are the *context* for each question and candidate algorithms are the *arms* to pull. We employ an EpochGreedy [4] type contextual bandits algorithm to provide a solution to the problem. The algorithm, when presented with a question, chooses a candidate ASAG technique to evaluate it and receives as feedback (from human expert/crowd), the best technique to evaluate the chosen question. Using this feedback, it updates its model of mapping questions to ASAG techniques and learns the best model for any given question over time. The algorithm is known to have excellent regret minimization guarantees [4] and we provide empirical evidence to show why it is suitable for the ASAG context.

2 Motivation

In a typical ASAG task, a model and a bunch of student answers are provided for a set of questions to predict students' scores. Additionally instructor evaluated scores (*groudtruth*) are also given against which the ASAG technique's scores (*predicted*) are evaluated. Let $\mathbf{u} = [u_1, u_2, \ldots, u_n]$ and $\mathbf{v} = [v_1, v_2, \ldots, v_n]$ denote the instructor given groundtruth scores and predicted scores by an ASAG technique for a question. The job of the evaluation measures is to determine how well $\mathbf{u}$ is represented by $\mathbf{v}$. In prior art, we find evidences of $\mathbf{u}$ and $\mathbf{v}$ to be of different data types with respect to the nomenclature proposed by Stevens [5]. **Nominal** type refers to discrete values of u_i (and v_j), not all of which are ordered [6]. On the other hand, both **ordinal** and **ratio** data provide ordering; with difference being in the former the gaps between successive values are not known (or uniform). Letter grades (A−, A, A+, B−, ...) are examples of ordinal data whereas scores of a question (0, 0.5, 1.5, 2) are examples of ratio data.

We observe that the distributions of $\mathbf{u}$ and $\mathbf{v}$ vary significantly across questions owing to their differences with respect to factors such as difficulty levels and scoring schemes. For example, Fig. 1 shows the distributions of groundtruth scores of the two questions shown in Table 1. Q-4 appears to be an easy one with most students getting the perfect 5/5 score compared to Q-2 which has a more spread-out scores over the entire range. Additionally, it appears to be more feasible to answer Q-2 in a partially

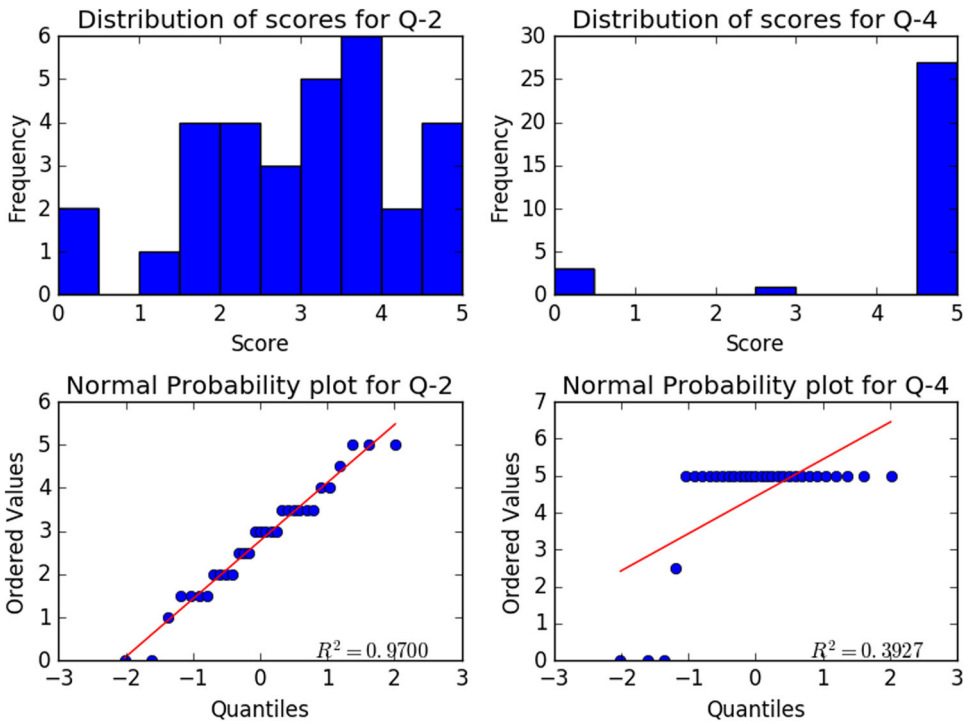

Fig. 1 Top row: frequency distribution of scores for two questions in CSD [3]. Bottom row: corresponding normal probability plots

correct manner, whereas Q-4 is of hit-or-miss nature. From the corresponding normal probability plots, it is evident that scores of Q-2 is significantly closer to a normal distribution than Q-4. Data being normally distributed and absence of outliers are two of the important prerequisites for application of Pearson's correlation coefficient (r) as an evaluation measure unlike some others such as MAE. Hence, the best technique for a question according to these two evaluation measures may not be always the same. Building on this intuition, Fig. 2 shows the best among five unsupervised ASAG techniques (described in Sect. 4.1) as per a range of evaluation measures (defined in Table 2) found in the ASAG literature. Depending on the desired evaluation measure, which is known only to the human expert (instructor), different techniques may emerge as the best for different questions.

In this paper, our key idea is to learn a mapping from a set of questions, represented using features from the question, model and student answers, to a set of ASAG techniques based on human feedback. After seeing sufficient number of example questions represented using meaningful features it is expected that a suitable algorithm, for example, EpochGreedy algorithm, will be able to predict the best ASAG technique when presented with a question.

3 Related work

In this section, we provide an overview of prior related research work. We start off with a summary view of techniques and evaluation measures used in the ASAG literature. Secondly, we present selected prior work in comparing different evaluation measures from various fields. Finally, we briefly refer to the work done in recent time on the topic of contextual bandits.

Automatic short answer grading Two recent survey papers by Roy et al. [2] and Burrows et al. [1] provide

Fig. 2 The best performing unsupervised ASAG technique (among ERB(0), SP(1), JC(2), LSA(3) and W2V(4) described in Sect. 4.1) as per different evaluation measures (defined in Table 2)

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21
MAE	3	2	1	4	4	2	3	3	3	3	4	3	3	3	3	3	3	3	3	3	3
RMSE	3	2	1	4	4	2	3	3	3	3	4	3	3	3	3	3	3	4	3	3	3
MZE	3	3	1	1	3	1	2	3	4	3	1	2	2	4	4	1	3	3	3	3	3
Pearson	4	2	1	4	4	2	3	4	3	3	4	2	3	3	3	2	4	2	1	3	2
Spearman	4	2	1	4	4	2	3	4	3	3	4	2	3	3	3	2	4	2	1	3	2
Kendall	4	2	3	4	3	2	3	4	4	3	4	1	3	3	3	1	1	1	2	3	4
Kappa	3	3	4	1	1	1	2	3	4	3	1	1	2	4	4	1	3	1	3	3	3
F1	3	3	4	1	1	4	2	3	1	3	4	1	2	4	4	1	4	3	4	3	3

Table 2 Commonly used evaluation measures for ASAG

Measure	Formula	Details
Mean absolute error ($MAE(\mathbf{u},\mathbf{v})$)	$\frac{1}{n}\sum_{i=1}^{n}\lvert u_i - v_i\rvert$	Mean absolute differences between elements of $\mathbf{u}$ and $\mathbf{v}$
Root mean square error ($RMSE(\mathbf{u},\mathbf{v})$)	$\{\frac{1}{n}\sum_{i=1}^{n}(u_i - v_i)^2\}^{\frac{1}{2}}$	Known to penalize large errors more heavily than MAE
Mean zero error ($MZE(\mathbf{u},\mathbf{v})$)	$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{u_i \neq v_i\}$	$\mathbb{1}\{.\}$ is an indicator variable
Pearson's $r(\mathbf{u},\mathbf{v})$	$\frac{\sum_{i=1}^{n}(u_i-\bar{u})(v_i-\bar{v})}{\sigma_u\sigma_v}$	$\bar{u}(\bar{v})$ is mean of $\mathbf{u}(\mathbf{v})$ and σ_u (σ_v) denotes SD of $\mathbf{u}(\mathbf{v})$
Spearman's $\rho(\mathbf{u},\mathbf{v})$	$1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$	d_i is the difference in rank of u_i and v_i i.e. where $\lvert\mathcal{R}(u_i,\mathbf{u}) - \mathcal{R}(v_i,\mathbf{v})\rvert$ and $\mathcal{R}(.,.)$ returns the rank of the first argument in the second
Kendall's $\tau(\mathbf{u},\mathbf{v})$	$\frac{n_c - n_d}{n(n-1)/2}$	$n_c(n_d)$ is the number of *concordant* (*discordant*) pairs. A pair (u_i, v_i) and (u_j, v_j) is said to be concordant (discordant) if both $u_i > u_j$ and $v_i > v_j$ or both $u_i < u_j$ and $v_i < v_j$ (if $u_i > u_j$ then $v_i < v_j$ or if $u_i < u_j$ then $v_i > v_j$)
Cohen's $\kappa(\mathbf{u},\mathbf{v})$	$\frac{p_o - p_e}{1 - p_e}$	p_o is the relative observed agreement, and p_e is the hypothetical probability of chance agreement
Precision (P($(\mathbf{u},\mathbf{v})$)	$\frac{1}{k}\sum_{j=1}^{k}\frac{\sum_{i=1}^{n}\mathbb{1}\{u_i=c_j \wedge v_i=c_j\}}{\sum_{i=1}^{n}\mathbb{1}\{v_i=c_j\}}$	
Recall(R($\mathbf{u},\mathbf{v}$))	$\frac{1}{k}\sum_{j=1}^{k}\frac{\sum_{i=1}^{n}\mathbb{1}\{u_i=c_j \wedge v_i=c_j\}}{\sum_{i=1}^{n}\mathbb{1}\{u_i=c_j\}}$	
F-measure (F($\mathbf{u},\mathbf{v}$))	$\{\alpha\frac{1}{P(\mathbf{u},\mathbf{v})} + (1-\alpha)\frac{1}{R(\mathbf{u},\mathbf{v})}\}^{-1}$	The balanced F measure (known as F_1) considers $\alpha = \frac{1}{2}$ and is the harmonic mean of precision and recall i.e. $F_1 = \frac{2PR}{P+R}$

$\mathbf{u}$ and $\mathbf{v}$ are as defined in Sect. 2

comprehensive views of research in ASAG. Both papers have grouped prior research based on the types of approaches used as well as the extent of human supervision needed. Broad categories include concept mapping [7], information extraction [8], supervised [9] and unsupervised [3] techniques. Both survey papers noted that a variety of evaluation measures have been used in ASAG literature. Broadly they can be grouped as *absolute error measures* such as mean absolute error (MAE), root mean square error (RMSE); various *correlation coefficients* such as the ones due to Pearson (r) and Spearman (ρ); as well as *confusion matrix based measures* such as Cohen's κ, precision, recall, and F-measures. We note that the choice of evaluation measure was influenced by the nature of the data and scoring scheme (nominal, ordinal or ratio scores) as well as type of techniques used (supervised, unsupervised) but remained greatly arbitrary. In this paper, we argue that irrespective of evaluation measure chosen, one can identify the best ASAG technique for a question over time with minimal feedback from human expert.

Evaluation measures In general, there have been plethora of research towards relative comparison of different evaluation measures for different tasks. Correlation measures have been most commonly used for evaluation of models. Several studies have compared Pearson's r with the related measure of Spearman's rank correlation coefficient (ρ) [10, 11]. Kendall's τ has been reported to have better statistical properties and offer a direct interpretation in terms of probabilities of observing concordant and discordant pairs [10, 12]. Powers et al. [13] compared the behavior of a number of evaluation measures, with emphasis on variations of κ-statistic under certain assumptions by simulating variations of data distributions.

Wilmott [14] debunked correlation measures in favor of absolute value measures such as MAE and RMSE. In a later paper, Wilmott and Matsuura [15] demonstrated MAE to be a superior measure than RMSE in assessing average model performance for classification tasks. They argued that RMSE, unlike MAE, varies with the variability within the distribution of error magnitudes and with the square root of the number of elements ($n^{\frac{1}{2}}$). Recently, Chai and Draxler [16] contradicted by asserting that RMSE is more appropriate to represent model performance than the MAE when the error distribution is expected to be Gaussian. Multiple studies have empirically compared various confusion matrix based evaluation measures for classification towards invariance, interpretability of the measures [17–19]. Cardoso and Sousa [20] brought out the pitfall of using traditional classification evaluation measures for evaluating ordinal regression tasks. They compared multiple measures and proposed a new measure emphasizing how much the predicted scores diverge from groundtruth and how inconsistent the predictions are with regard to the relative order of the classes.

Contextual bandits The contextual bandits setting is an extension of the standard multi-armed bandits (MAB) setting which has been studied since [21]. Optimal algorithms (in terms of regret) for the standard MAB setting

have been developed more recently both in the stochastic [22] as well as adversarial [23] settings. The problem of contextual bandits has been studied as MABs with side information. Auer et al. [23] proposed the EXP4 algorithm for this setting. Langford and Zhang [4] proposed the EPOCHGREEDY algorithm more recently for the same setting which has better regret guarantees under certain conditions. Several variations and extensions of EPOCH-GREEDY have been proposed to make it more scalable and computationally faster [24]. While all the algorithms proposed have focused extensively on proving theoretical guarantees for the contextual bandits setting, fewer works have focused on applying it in practice. Hofmann, Whitson and de Rijke [25] applied contextual bandits framework for information retrieval problems while recently Lan and Baraniuk [26] used it to recommend personalized learning actions to students to maximize their success.

4 Techniques and measures for ASAG

In this section, we provide brief descriptions of ASAG techniques under consideration and commonly used evaluation measures in the literature.

4.1 Techniques

Most prior research in ASAG were done on proprietary datasets and are not available in public domain. Secondly, we observe prevalence of dataset specific feature engineering. Hence, we implemented representative unsupervised and supervised ASAG techniques as candidate algorithms for our empirical study.

4.2 Unsupervised techniques

These ASAG techniques leverage various textual similarity measures (lexical, semantic etc.) to obtain a similarity value between student and model answers and convert those values to grades appropriately. They largely reduce the need of instructor involvement either for training the ASAG system as done in their supervised counterparts. We implemented representative techniques which were used in prior ASAG research as well as recently popular distributional semantics based similarity measure [27]. These techniques output a continuous value score in [0, 1] which are scaled and discretized for *absolute error measures* and *confusion matrix based measures*.

- *ERB* ERB or Evaluating Responses with BLEU is a lexical measure comparing student answers against model answers using a modified version of the n-gram co-occurrence scoring algorithm called BLEU [28], commonly used for machine translation evaluation [29]. We tried different n between 1 and 5 though owing to short length of answers longer n-grams do not get enough support.
- *JC and SP* These are two semantic similarity measures based on Wordnet [30]. For each word in student answer, maximum word-to-word similarity scores are obtained with respect to words in model answers which are then summed up and normalized by the length of the two responses as described by Mohler and Mihalcea [3]. They compared eight options for computing word-to-word similarities; of which we select the two best performing ones viz. the measure proposed by Jiang and Conrath (**JC**) [31] and Shortest Path (**SP**).
- *LSA and W2V* These are the measures in vector space similarity category. In this category we first chose the most popular measure for measuring semantic similarity viz. Latent Semantic Analysis (**LSA**) [32] trained on a Wikipedia dump. We also use the recently popular word2vec tool (**W2V**) [27] to obtain vector representation of words which are trained on 100 billion words of Google news dataset and are of length 300. Word-to-word similarity measures obtained using euclidean distance between word vectors are summed up and normalized in a manner similar to *JC* and *SP*.

4.3 Supervised techniques

This group of techniques needed instructors to grade a fraction of student answers (typically ranging from half to three-quarters) to train *supervised learning algorithms* as training data for building classification or regression models. We consider discrete class labels and hence model this as a classification problem and use support vector machine (SVM). We use unsupervised similarity measures described above and a few dataset specific features for both.

4.4 Transfer learning

Roy et al recently demonstrated an interesting technique for ASAG using transfer learning [33]. They proposed an iterative technique using an ensemble of numeric and text classifiers where they brought down labeling requirement by using graded answers from one (source) question to score answers to another (target) question.

4.5 Performance measures

Table 2 shows the commonly used measures which we will refer to for presenting empirical results. While we find at least one evidence of each of these measures in prior

literature, some of these such as Pearson's r, MAE, RMSE and Cohen's κ are found more than the rest.

5 Contextual bandits based algorithm

Algorithm 1 EPOCHGREEDY

Initialize $W_0 = \{\}, t_1 = 1$
for $\ell = 1, 2 \ldots$ **do**
  $t = t_\ell$
  Receive question $q^t \in \mathcal{Q}$
  Choose algorithm $a^t \in \mathcal{A}$ uniformly at random
  Receive reward $r^t(q^t, a^t) \in \mathbb{R}_+$
  $W_\ell = W_\ell \cup \{(q^t, a^t, r^t)\}$
  Find best hypothesis h_ℓ by solving

$$h_\ell = \arg\max_{h \in \mathcal{H}} \sum_{(q,a,r) \in W_\ell} r\mathbb{I}(h(q) = a)$$

  $t_{\ell+1} = t_\ell + s(W_\ell) + 1$
  for $t = t_\ell + 1, \ldots, t_{\ell+1} - 1$ **do**
    Select arm $a^t = h_\ell(q^t)$
    Receive reward $r^t(q^t, a^t) \in \mathbb{R}_+$
  end for
end for

In this section, we first briefly describe the multi-armed bandit problem and the formulation of selection of ASAG techniques using the contextual variation. The multi-armed bandit (MAB) problem attempts to learn a policy for a player that maximizes the total expected reward by playing a collection of slot machines. At each turn the player pulls an arm in the slot machine for a fixed number of turns. Each machine has a fixed reward distribution which the player does not have any prior information about. The player conducts trials in two mode viz. *exploration* (trying out arms which might yield high rewards) and *exploitation* (playing arms which has given the highest observed reward till then). The objective is to maximize the total reward over a long period of time. Contextual variations of MABs extend the MAB framework by incorporating additional information, known as the context or side-information. The context could be available about the player and/or the machines with the objective better and/or faster arrival at the optimal choice of the policy.

We formulate the problem of selecting evaluation measures for ASAG techniques using contextual MAB framework. In an analogous manner to the MAB setting, an ASAG system must strike a balance between testing the efficacy of different candidate algorithms (exploration) and maximizing the accuracy measured by various evaluation measures using observations on the actions (exploitation). Contexts in our case come in the form of various features of questions and instructor-provided model answers. The features are designed in a manner so that they characterize various linguistic aspects of questions, model answers and expected student answers.

Next we describe the EPOCHGREEDY algorithm of [4] for the contextual bandits problem. The algorithm, as the name suggests, proceeds in epochs. Specifically the algorithm has two types of epochs namely *exploration epoch* and *exploitation* epoch. In the exploration epoch (which is just one iteration), the algorithm chooses arms uniformly at random and observes rewards for the chosen arm. At the end of the exploration epoch, the algorithm performs an empirical risk minimization over a hypothesis space $\mathcal{H}$ to choose a hypothesis h (mapping from questions to techniques) that best predicts the rewards seen so far. Then the exploitation epoch begins where the algorithm plays the arm suggested by the hypothesis learnt at the end of the exploration epoch. The algorithm is shown in Algorithm 1

The regret of the EPOCHGREEDY (EG) algorithm after T rounds is defined as

$$R_T(\mathrm{EG}) = \sum_{\mathrm{T}=1}^{\mathrm{T}} \mathrm{R}^{\mathrm{T}}\left(\mathrm{Q}^{\mathrm{T}}, \mathrm{EG}(\mathrm{Q}^{\mathrm{T}})\right) - \arg\min_{\pi \in \Pi} \sum_{\mathrm{T}=1}^{\mathrm{T}} \mathrm{R}^{\mathrm{T}}\left(\mathrm{Q}^{\mathrm{T}}, \pi(\mathrm{Q}^{\mathrm{T}})\right)$$

where $\Pi : \mathcal{Q} \to \mathcal{A}$ is the space of comparator mappings from questions to techniques. By a suitable choice of the epoch size $s(W_\ell)$, the epoch greedy algorithm can be shown to have a regret that scales as $O(T^{2/3}(K \ln(m))^{1/3})$. Here, K is the number of techniques and m is the number of hypothesis in the class $\mathcal{H}$. When considering classes with infinite hypotheses (such as linear separators), the m can be replaced with the VC-dimension of the comparator class $\mathcal{H}$. Further analysis of the regret under additional assumptions can be found in [4].

6 Empirical evaluation

For empirical validation, we use one of the earliest ASAG datasets comprising of a set of questions, model answers and student answers taken from an undergraduate computer science course [3] (CSD). The data set consists of 21 questions (7 questions from 3 assignments each) from introductory assignments in the course with answers provided by a class of about 30 undergraduate students. Student answers were independently evaluated by two annotators on a scale of 0-5 and automatic techniques are measured against their average.

As the number of questions and number of students are relatively small in available datasets, we conducted simulations to apply EPOCHGREEDY on CSD. We choose multi-class Support Vector Machine (SVM) with Gaussian kernel for empirical risk minimization during exploration based on received rewards (r^t) and select a technique (a^t) during exploitation. The SVM uses the following features as

context to represent questions (q^t). These features are extensible to other dataset specific features which can be computed without requiring groundtruth scores.

- Mean length of student answers.
- Standard deviation of length of student answers.
- Lexical diversity of student answers in terms of type-token ratio [34].
- Mean and standard deviation of similarity between student answers with the corresponding model answer using different similarity measures described in Sect. 4.1.

Next we present a different simulation scenario and corresponding results in terms of regret values over time epochs. For all our experiments, we use *epoch size* to be 10 and all results are averaged over 100 simulation runs.

6.1 Sampling with replacement

During simulation, at every epoch, the learning algorithm chooses one of the actual questions and generates its context. The change in the regret against time epoch is shown in Fig. 3 for all the evaluation measures. Given that the number of questions is relatively small, the classifier within a few intervals managed to see and learn from all the questions. Consequently, regret tends to zero within a small number of epochs for all evaluation measures; though for some measures such as Pearson's r and F1 the initial regrets are quite high.

6.2 Similar questions

In a real-world setting, the new questions are not expected to be exactly same but similar at best to one or more of the past questions. In this setting, during simulation, questions' features are perturbed by adding some random noise to their values to mimic the scenario of such similar questions. We vary the noise amount in a range [0, 0.5] at discrete step of 0.1 and plot regret against epoch in Fig. 4. While the regret becomes small over time even at a high noise level, the time taken (number of epochs) increases expectedly with noise level.

6.3 Noisy human experts

This setting simulates the scenario where the human expert is fallible i.e. some of the reward obtained are incorrect. Such noise can come from two sources viz. (1) instead of an expert, we take feedback from a crowd or (2) the human expert looks at only a fraction of student answers while giving feedback to EPOCHGREEDY and thereby providing noisy reward. Figure 5 shows the number of epochs taken for the regret to converge to a small value against noise in human feedback. With increase in human noise the rate of convergence becomes slower for evaluation measures though for some of the measures (e.g. MAE), the rate is not monotonically increasing.

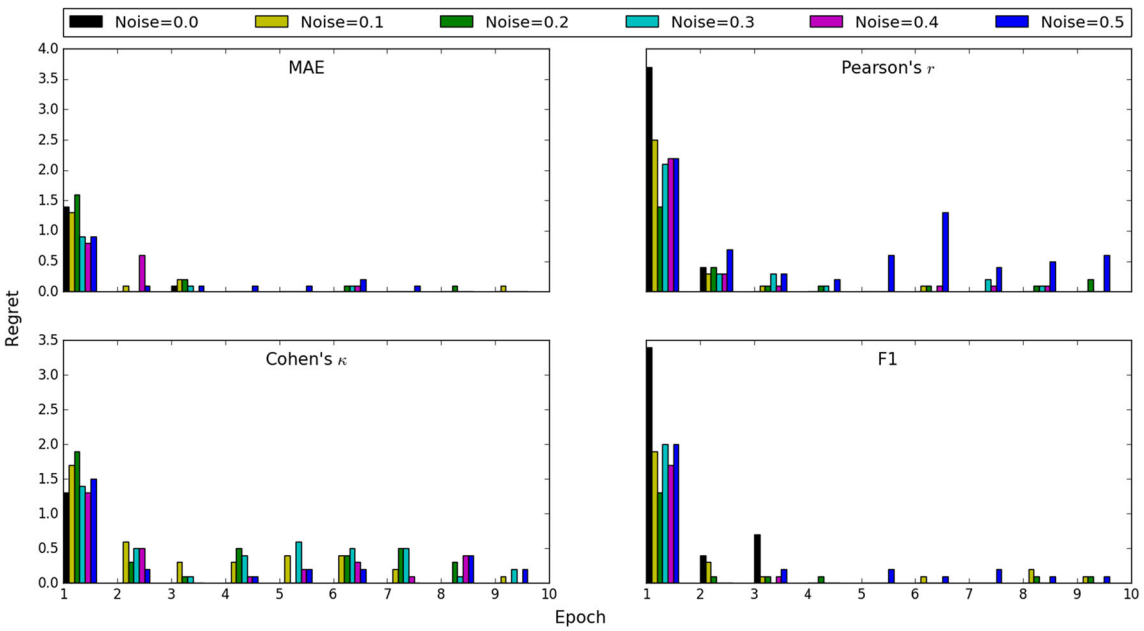

Fig. 3 Sampling with replacement: for all evaluation measures the regret becomes small within a few time epochs for CSD

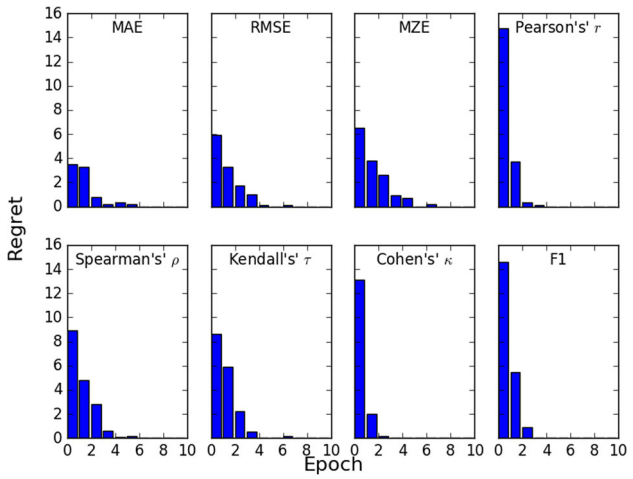

Fig. 4 Feature noise: at higher noise level it takes longer for the regret to become small (best viewed in color) (color figure online)

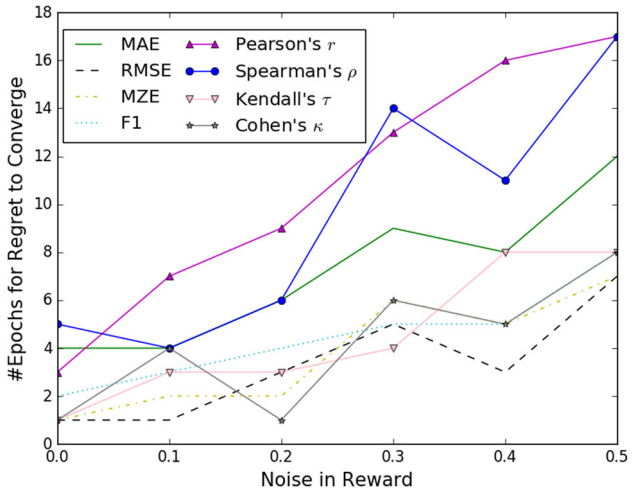

Fig. 5 Label noise: with increase of noise in reward received, number of epochs to converge increases (best viewed in color) (color figure online)

7 Conclusion

In this paper, we analysed the role of evaluation measures used for assessing the quality of automatic short answer grading techniques. We demonstrated that factors such as difficulty level and diversity of student answers vary significantly across questions leading to different ASAG techniques emerging as superior for different evaluation measures. Building on this observation, we proposed to automatically *learn* a mapping from questions to ASAG techniques using minimal human (expert/crowd) feedback. We formulated this learning task as a contextual bandits problem and provided a rigorous regret minimization algorithm that handles key practical considerations such as noisy experts and similarity between questions. We have presented extensive simulations on a standard dataset to demonstrate that our approach provides outcomes which are consistent with human evaluations. In future, we intend to consider larger real-world datasets with a large number of questions, e.g. arising from a massive online open course (MOOC) setting, towards assessing applicability and efficacy of the proposed technique.

References

1. Burrows, S., Gurevych, I., Stein, B.: The eras and trends of automatic short answer grading. Int. J. Artif. Intell. Educ. **25**(1), 60–117 (2015)
2. Roy, S., Narahari, Y., Deshmukh, O.D.: A perspective on computer assisted assessment techniques for short free-text answers. In: Computer Assisted Assessment. Research into E-Assessment, pp. 96–109. Springer (2015)
3. Mohler, M., Mihalcea, R.: Text-to-text semantic similarity for automatic short answer grading. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL). Association for Computational Linguistics (2009)
4. Langford, J., Zhang, T.: The epoch-greedy algorithm for multi-armed bandits with side information. In: Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Dec 3–6, 2007, pp. 817–824 (2007)
5. Stevens, S.S.: On the theory of scales of measurement (1946)
6. Myroslava, O., et al.: Semeval-2013 task 7: the joint student response analysis and 8th recognizing textual entailment challenge. Technical report (2013)
7. Leacock, C., Chodorow, M.: C-rater: automated scoring of short-answer questions. Comput. Humanit. **37**(4), 389–405 (2003)
8. Mitchell, T., Russell, T., Broomhead, P., Aldridge, N.: Towards robust computerized marking of free-text responses. In: Proceedings of 6th International Computer Aided Assessment Conference (2002)
9. Madnani, N., Burstein, J., Sabatini, J., OReilly, T.: Automated scoring of a summary writing task designed to measure reading comprehension. In: Proceedings of the 8th Workshop on Use of NLP for Building Educational Applications (2013)
10. Hauke, J., Kossowski, T.: Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. Quaest. Geogr. **30**(2), 87–93 (2011)
11. Mukaka, M.M.: A guide to appropriate use of correlation coefficient in medical research. Malawi Med. J. **24**(3), 69–71 (2012)
12. Newson, R.: Parameters behind "nonparametric" statistics: Kendall's tau, somers' d and median differences. Stata J. **2**(1), 45–64(20) (2002)
13. Powers, D.M.W.: The problem with kappa. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL), pp. 345–355. The Association for Computer Linguistics (2012)
14. Willmott, C.J.: Some comments on the evaluation of model performance. Bull. Am. Meteorol. Soc. **63**, 1309–1369 (1982)
15. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. Clim. Res. **30**(1), 79–82 (2005)
16. Chai, T., Draxler, R.R.: Root mean square error (rmse) or mean absolute error (mae)?—arguments against avoiding rmse in the literature. Geosci. Model Dev. **7**(3), 1247–1250 (2014)
17. Ferri, C., Hernndez-Orallo, J., Modroiu, R.: An experimental comparison of performance measures for classification. Pattern Recognit. Lett. **30**(1), 27–38 (2009)

18. Freitas, A.A.: Comprehensible classification models: a position paper. ACM SIGKDD Explor. Newslett. **15**(1), 1–10 (2014)
19. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. **45**(4), 427–437 (2009)
20. Cardoso, J.S., Sousa, R.G.: Measuring the performance of ordinal classification. Int. J. Pattern Recognit. Artif. Intell. **25**(8), 1173–1195 (2011)
21. Robbins, H.: Some aspects of the sequential design of experiments. Bull. Am. Math. Soc. **58**(5), 527–535 (1952)
22. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. Mach. Learn. **47**(2–3), 235–256 (2002)
23. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: The nonstochastic multiarmed bandit problem. SIAM J. Comput. **32**(1), 48–77 (2002)
24. Agarwal, A., Hsu, D.J., Kale, S., Langford, J., Li, L., Schapire, R.E.: Taming the monster: a fast and simple algorithm for contextual bandits. In: CoRR, abs/1402.0555 (2014)
25. Hofmann, K., Whiteson, S., de Rijke, M.: Contextual bandits for information retrieval. In: NIPS 2011 Workshop on Bayesian Optimization, Experimental Design, and Bandits, vol. 12 (2011)
26. Lan, A.S., Baraniuk, R.G.: A contextual bandits framework for personalized learning action selection. In: Proceedings of the 9th International Conference on Educational Data Mining, pp. 424–429 (2016)
27. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
28. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. Technical report, IBM Research Report (2001)
29. Perez, D., Alfonseca, E., Rodríguez, P.: Application of the bleu method for evaluating free-text answers in an e-learning environment. In: LREC, European Language Resources Association (2004)
30. Miller, G.A.: Wordnet: a lexical database for english. Commun. ACM **38**(11), 39–41 (1995)
31. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the International Conference on Research in Computational Linguistics (1997)
32. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Process. **25**(2–3), 259–284 (1998)
33. Roy, S., Bhatt, H.S., Narahari, Y.: Transfer learning for automatic short answer grading. In: Proceedings of the European Conference on Artificial Intelligence (ECAI) (2016)
34. Lieven, E.V.M.: Conversations between mothers and young children: individual differences and their possible implication for the study of language learning. na (1978)