

How to Take Into Account a Student's Degree of Certainty When Evaluating the Test Results

Joe Lorkowski¹, Olga Kosheleva², and Vladik Kreinovich¹

Departments of ¹Computer Science and ²Teacher Education

University of Texas at El Paso, El Paso, TX 79968, USA

lorkowski@computer.org, olgak@utep.edu, vladik@utep.edu

Abstract—To more adequately gauge the student's knowledge, it is desirable to take into account not only whether the student's answers on the test are correct or nor, but also how confident the students are in their answers. For example, a situation when a student gives a wrong answer, but understands his/her lack of knowledge on this topic, is not as harmful as the situation when the student is absolutely confident in his/her wrong answer. In this paper, we use the general decision making theory to describe the best way to take into account the student's degree of certainty when evaluating the test results.

I. INTRODUCTION

Need to take into account the student's degree of certainty.

On a usual test, a student provides answers to several questions, and the resulting grade depends on whether these answers are correct.

However, this approach does not take into account how confident the students is in his/her answer.

In real life situations, when a person needs to make a decision, it may happen that a person does not know the correct answer – but it is not so bad if this person realizes that his knowledge is weak on this subject. In this case, he/she may consult an expert and come up with a correct decision. The situation is much worse if a decision maker is absolutely confident in his/her wrong decision.

For example, we do not expect our medical doctor to be perfect, but what we do expect is that when the doctor is not sure, he/she knows that his/her knowledge of this particular medical situation is lacking, and either consults a specialist him/herself or advises the patient to consult a specialist.

From this viewpoint, when gauging the student's level of knowledge, it is desirable:

- to explicit ask the student how confident he/she is in the corresponding answer, and
- to take this degree of confidence into account when evaluating the resulting grade.

Some tests already solicit these confidence degrees from the students; see, e.g., [4], [6], [8], [9]; see also [3], [8].

How can we take the student's degree of certainty into account? Once we have the student's degrees of confidence in different answers, how should we combine these degrees of confidence into a single overall grade?

As of now, the existing combination rules are semi-heuristic. It is therefore desirable to come up with well-justified rules for such combination.

What we do in this paper. In this paper, we propose to use decision making theory to come up with a combination rule that adequately describes the effect of possible uncertainty and/or wrong answers on the possible decisions.

For that, we emulate a real-life decision making situation, when the decision is made by a group of specialists including the current student. In this setting, we estimate the expected contribution of this student's knowledge to the quality of the resulting decision.

II. HOW TO TAKE INTO ACCOUNT A STUDENT'S DEGREE OF CERTAINTY WHEN EVALUATING THE TEST RESULTS: CASE OF ONE QUESTION WITH TWO ALTERNATIVES

Need to describe how decisions are made: reminder. A natural idea to gauge student's uncertain knowledge is to analyze how this uncertainty affects the decisions. To be able to perform this analysis, we need to describe what is a reasonable way to make a decision based on the opinion of several uncertain (and independent) experts.

Decision making under uncertainty: general case. According to decision theory, decisions made by a rational agent can be described as follows:

- to each possible situation, we assign a numerical value called its *utility*, and
- we select an action for which the expected value of utility is the largest possible;

see, e.g., [1], [5], [7], [10].

Let us start with the simplest simplified case. Let us start our analysis with the simplest case, in which:

- we only have one question and
- for this question, there are only two possible alternatives A_1 and A_2 .

Let P_1 be the student's degree of confidence in the answer P_1 . Since we assumed that there are only two possible answers, the student's degree of confidence in the other answer A_2 is equal to $P_2 = 1 - P_1$.

To gauge the effect of the student's answer on the resulting decision, let us assume that for each of the two alternatives A_1 and A_2 , we know the optimal action.

For example, we have two possible medical diagnoses, and for each of these diagnoses, we know an optimal treatment.

Let $u_{i,j}$ be the utility corresponding to the case when

- the actual situation is A_i and
- we use the action which is optimal for the alternative A_j .

In these terms, the fact that the action corresponding to A_1 is optimal for the situation A_1 means that $u_{1,1} > u_{1,2}$; similarly, we get $u_{2,2} > u_{2,1}$.

If we know the probabilities p_1 and $p_2 = 1 - p_1$ of both situations, then we select the action corresponding to A_1 if its expected utility is larger, i.e., if

$$p_1 \cdot u_{1,1} + (1 - p_1) \cdot u_{2,1} \geq p_1 \cdot u_{1,2} + (1 - p_1) \cdot u_{2,2},$$

i.e., equivalently, if

$$p_1 \geq t \stackrel{\text{def}}{=} \frac{u_{2,2} - u_{2,1}}{(u_{1,1} - u_{1,2}) + (u_{2,2} - u_{2,1})}. \quad (1)$$

If the actual situation is A_1 , then the optimal action is the one corresponding to A_1 . Thus, the above inequality describes when the optimal action will be applied – when our degree of confidence in A_1 exceeds a certain threshold t .

How to estimate the probabilities of different alternatives under expert uncertainty. Let us denote the number of experts by n . Let us assume that for each expert k , we know this expert's degree of confidence (subjective probability) $p_{1,k}$ in alternative A_1 , and his/her degree of confidence $p_{2,k} = 1 - p_{1,k}$ in alternative A_2 .

In general, we do not have prior reasons to believe that some experts are more knowledgeable than others, so we assume that all n experts are equally probable to be right:

$$P(k\text{-th expert is right}) = \frac{1}{n}.$$

Thus, by the law of complete probability, we have

$$p_1 = \text{Prob}(A_1 \text{ is the actual alternative}) =$$

$$\sum_{k=1}^n P(k\text{-th expert is right}) \cdot P(A_1 | k\text{-th expert is right}),$$

hence

$$p_1 = \frac{1}{n} \cdot \sum_{k=1}^n p_{1,k}. \quad (2)$$

How to estimate the student's contribution to the correct decision. We started with the average of the probabilities of n experts. Once we add the student as a new expert, with probabilities $p_{1,n+1} = P_1$ and $p_{2,n+1} = P_2$, the probability p_1 changes to the new value

$$p'_1 = \frac{1}{n+1} \cdot \sum_{k=1}^{n+1} p_{1,k} = \frac{n}{n+1} \cdot p_1 + \frac{1}{n+1} \cdot P_1. \quad (3)$$

For large n , this addition is small, so in most cases, it does not change the decision. However:

- sometimes, the increase in the estimated probability p_1 will help us switch from the wrong decision to the correct one, and
- sometimes, vice versa, the new estimate will be smaller than the original one and thus, because of the addition of the student's opinion, the group will switch from the correct to the wrong decision.

According to the general decision theory ideas, the student's contribution can be gauged as the expected utility caused by the corresponding change, i.e., as the probability of the positive change times the gain minus the probability of the negative loss times the corresponding loss.

The probability of a gain is equal to the probability that $p_1 < t$ but $p'_1 \geq t$. Due to (3), the inequality $p'_1 \geq t$ is equivalent to

$$p_1 \geq t + \frac{1}{n} \cdot t - \frac{1}{n} \cdot P_1.$$

Thus, the probability of the gain is equal to the probability that the previous estimate p_1 is in the interval

$$\left[t + \frac{1}{n} \cdot t - \frac{1}{n} \cdot P_1, t \right].$$

For large n , this interval is narrow, so this probability can be estimated as the probability density $\rho(t)$ of the probability corresponding to p_1 times the width

$$\frac{1}{n} \cdot P_1 - \frac{1}{n} \cdot t$$

of this interval. Thus, this probability is a linear function of P_1 .

Similarly, the probability of the loss is also a linear function of P_1 and hence, the expected utility also linearly depends on P_1 .

How to gauge student's knowledge: analysis of the problem. The appropriate measure should be a linear function of the student's degree of certainty P_1 . So, if we originally assign:

- N points to a student who is absolutely confident in the correct answer and
- 0 points to a student who is absolutely confident in the wrong answer,

then the number of points assigned in general should be a linear function of P_1 that is:

- equal to N when $P_1 = 1$, and
- equal to 0 when $P_1 = 0$.

One can check that the only linear function with this property is the function $N \cdot P_1$. Thus, we arrive at the following recommendation:

How to gauge student's knowledge: the resulting recommendation. When a student supplies his/her degree of confidence P_1 in the answer A_1 (and, correspondingly, the degree of confidence $P_2 = 1 - P_1$ in the answer A_2), then we should give the student:

- $N \cdot P_1$ points if A_1 is the correct answer, and
- $N \cdot P_2$ points if A_2 is the correct answer,

where N is the number of points that the student would get for an absolutely correct answer with confidence 1.

Discussion. Let us show that if we follow the above recommendation, then we assign different numbers of grades in two situations that we wanted to distinguish:

- the bad situation in which a student is absolutely confident in the wrong answer $p_1 = 0$ and $p_2 = 1$, and
- a not-so-bad situation when a student is ignorant but understands his or her ignorance and assigns the degree of confidences $p_1 = p_2 = 0.5$ to both possible answers.

Indeed:

- In the first (bad) situation, the student gets the smallest number of points: 0.
- In the second, not-so-bad situation, the student gets $N \cdot 0.5$ points, which is more than 0.

III. GENERAL CASE

What if a question has several possible answers: analysis of the problem. In general, a question can have several possible answers corresponding to several alternatives. Let us denote these alternatives by A_1, \dots, A_s .

In this case, a student assigns, to each of these alternatives A_i , his/her degree of certainty P_i . Since we know that exactly one of the given s alternatives is true, these probabilities should add up to 1: $\sum_{i=1}^s P_i = 1$.

How does adding this student's knowledge change the decision of n experts? Similarly to the previous case:

- it may be that previously, the experts selected a wrong alternative, and the student's knowledge can help select the correct alternative A_1 ;
- it also may be that the experts selected the correct alternative, but the addition of the student's statement will lead to the selection of a wrong alternative.

Let us describe this in detail.

Similarly to the case of two alternative, we can conclude that a group of experts selects an action corresponding to the alternative A_{i_0} if the corresponding expected utility is larger than the expected utility of selecting any other action, i.e., if

$$\left(\sum_{j=1}^s p_j \cdot u_{i_0,j} \right) - \left(\sum_{j=1}^s p_j \cdot u_{i,j} \right) \geq 0 \quad (4)$$

for all i , where p_i is the estimate of the probability that the i -th alternative A_i is true. Similarly to the case of two alternatives, we conclude that

$$p_i = \frac{1}{n} \cdot \sum_{k=1}^n p_{i,k},$$

where n is the number of experts and $p_{i,k}$ is the estimate of the probability of the i -th alternative A_i made by the k -th expert.

When we add, to n original experts, the student as a new expert, with $p_{i,n+1} = P_i$, then the probabilities p_i change to new values

$$p'_i = \frac{1}{n+1} \cdot \sum_{k=1}^{n+1} p_{i,k} = \frac{n}{n+1} \cdot p_i + \frac{1}{n+1} \cdot P_i.$$

Thus, the left-hand side of the inequality (4) has a change which is linear in terms of the degrees P_1, \dots, P_s .

For each case when the addition of the student as a new expert changes the inequality between two expected utilities, the corresponding interval of possible values of the difference is small and thus, the resulting utility is proportional to the linear function of P_i – and is, thus, linear as well.

The probability of each such change is very small, so the probability that the addition of a student can change two or more inequalities – i.e., that two changes can occur at the same time – can be estimated as the product of these two (or more) small numbers and can, therefore, be safely ignored.

In this approximation, the overall utility can be obtained by adding the probabilities of all such cases and is, therefore, also a linear function of the probabilities P_i :

$$u = u_0 + \sum_{i=1}^s a_i \cdot P_i.$$

Let c be the index of the correct answer, then this formula can be reformulated as

$$u = u_0 + u_c \cdot P_c + \sum_{i \neq c} a_i \cdot P_i. \quad (5)$$

Usually, there are no a priori reasons why one incorrect answer is better than another incorrect answer. So, it is natural to assume that the utility c_i corresponding to each incorrect answer A_i is the same. Let us denote this common value of the utility by f . Then, the above formula (5) takes the form

$$u = u_0 + u_c \cdot P_c + f \cdot \sum_{i \neq c} P_i. \quad (6)$$

Since the degrees P_i add up to 1, we have

$$P_c + \sum_{i \neq c} P_i = 1,$$

hence

$$\sum_{i \neq c} P_i = 1 - P_c,$$

and the formula (6) takes the form

$$u = u_0 + u_c \cdot P_c + f \cdot (1 - P_c).$$

Thus, the utility is a linear function of the student's degree of confidence P_c in the correct answer.

Let us denote by N the number of points that we assign to a correct answer in which the student is fully confident ($P_c = 1$). Naturally, a student gets 0 points when he or she is fully confident in the wrong answer (i.e., $P_i = 1$ for some $i \neq c$ and thus, $P_c = 0$). Thus, the desired linear function

should be equal to N when $P_c = 1$ and to 0 when $P_c = 0$. There is only one such linear function: $N \cdot P_c$. So, we arrive at the following recommendation.

What if there are several possible answers: the resulting recommendation. Let A_1, \dots, A_s be possible answers, out of which only one answer A_c is correct. Let N denote the number of points that a student would get for a correct answer in which he or she is absolutely confident.

During the test, the student assigns, to each possible answer A_i , his/her degree of confidence P_i that this answer is correct. These degrees must add up to 1: $\sum_{i=1}^s P_i = 1$.

Our analysis shows that for this, we give the student $N \cdot P_c$ points, where P_c is the student's degree of confidence in the correct answer.

How do we combine grades corresponding to different problems? In the above text, we describe how to assign number of points to a *single* question. Namely:

- Our idea was to assign the number of points which is proportional to the gain in expected utility that the student's answer can bring in a real decision making situation.
- Our analysis has shown that this expected utility is proportional to the probability P_1 of the correct answer.
- Thus, our recommendation is to assign the number of points proportional to the probability of the correct answer.

Real-life tests usually have *several* questions.

- In the traditional setting, a student provides answers to each of these questions.
- In the new setting, for each question, the student also provides us with his/her degrees of confidence in different possible answers to this question.

How do we gauge the student knowledge level based on all this information?

A natural idea is – similarly to the case of a single question – to use, as the measure of the student's knowledge, the expected utility that the student's answers can bring in a real decision making situation. Let us show how this idea can be applied.

The general decision making situation means selecting a decision for each of the problems. For example, on a medical exam, a student may be asked several questions describing different patients.

Usually, different questions on the test are independent from each other. It is known (see, e.g., [2]) that if a decision problem consists of several independent decisions, then the utility of each combination of selections is equal to the sum of the corresponding utilities.

We know the utility corresponding to each question – this is the value that we used as a recommended grade for this particular question. Thus, the overall grade for the test should

be equal to the sum of the grades corresponding to individual questions.

Hence, we arrive at the following recommendation.

IV. RESULTING RECOMMENDATION

Let us consider a test with T questions $q_1, \dots, q_t, \dots, q_T$. For each question q_t , a student is given several possible answers $A_{t,1}, A_{t,2}, \dots$. For each question q_t , we know the number of points N_t corresponding to the answer which is correct and for which the student has a full confidence.

The student is required, for each question q_t and for each possible answer $A_{t,i}$, to provide his/her degree of confidence $P_{t,i}$ that this particular answer is correct. For each question q_t , these probabilities should add up to 1:

$$P_{t,1} + P_{t,2} + \dots = 1.$$

To estimate the student's level of knowledge, we need to know, for each question q_t , the correct answer; let us denote this correct answer by $A_{t,c(t)}$. Then:

- for each question q_t , we give the student $P_{t,c(t)} \cdot N_t$ points;
- as an overall grade g , i.e., as a measure of overall student knowledge, we take the sum of the points given for individual problems: $g = \sum_{t=1}^T P_{t,c(t)} \cdot N_t$.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721.

The authors are very thankful to Dr. Michelle D. Miller for her very interesting talk that inspired this research.

REFERENCES

- [1] P. C. Fishburn, *Utility Theory for Decision Making*, Wiley, New York, 1969.
- [2] P. C. Fishburn, *Nonlinear Preference and Utility Theory*, The John Hopkins Press, Baltimore, Maryland, 1988.
- [3] R. M. Isaacson and C. A. Was, "Building a metacognitive curriculum: An educational psychology to teach metacognition", *National Teaching & Learning Forum*, 2010, Vol. 19, No. 5, pp. 1–4.
- [4] M. Lovette, O. Meyer, and C. Thille, "The Open Learning Initiative: Measuring the effectiveness of the OLI statistics course in accelerating student learning", *Journal of Interactive Media in Education*, 2008, Vol. 1, pp. 1–16.
- [5] D. R. Luce and H. Raiffa, *Games and Decisions, Introduction and Critical Survey*, Wiley, New York, 1957.
- [6] M. D. Miller, *Minds Online: Teaching Effectively with Technology*, Harvard University Press, Cambridge, Massachusetts, 2014.
- [7] R. B. Myerson, *Game theory: Analysis of Conflict*, Harvard University Press, Cambridge, Massachusetts, 1991.
- [8] L. B. Nilson, *Creating Self-Regulated Learners: Strategies to Strengthen Students' Self-Awareness and Learning Skills*, Stylus Publishing, Sterling, Virginia, 20013.
- [9] Open Learning Initiative (OLI), Cargenie Mellon Unuversity, <http://oli.cmu.edu>
- [10] P. Suppes, D. M. Krantz, R. D. Luce, and A. Tversky, *Foundations of Measurement. Vol. II. Geometrical, Threshold, and Probabilistic Representations*, Academic Press, San Diego, California, 1989.