Feedback in Peer Assessment for Open-Response Assignments Using a Multitask Factorization Approach

Jorge Díez
Artificial Intelligence Center
Universidad de Oviedo
Gijon, Asturias, Spain
jdiez@uniovi.es

Oscar Luaces
Artificial Intelligence Center
Universidad de Oviedo
Gijon, Asturias, Spain
oluaces@uniovi.es

Antonio Bahamonde
Artificial Intelligence Center
Universidad de Oviedo
Gijon, Asturias, Spain
abahamonde@uniovi.es

Abstract

To leverage the benefits of assessments in the learning process students should receive some feedback that explains the reasons of a global grade. When peer assessment is involved, the output of the operation is typically just one grade that has a limited value in order to improve the knowledge of the students. In fact, this is one of the criticisms of peer assessment. In this paper we present a method to provide the students with an additional feedback after conducting an assessment of open-response assignments. The students are asked to evaluate a number of different aspects of the answers of other students. One of these aspects is the global grade, but there are other annotations that can be included to explain the global grade. We implemented the aspects to be assessed as labels with an ordinal level, and then the value of all these labels can be learned using a multitask approach. A consequence of this approach is that peer assessment can be extended to grade other answers not considered during the multitask training. Therefore, this method can reduce significantly the burden on students; another flaw of peer assessment. Finally, after presenting the method, we report a number of experiments carried out with 3 datasets obtained from courses of different fields of our university.

1 Introduction

The assessment of open-response assignments is frequently a problem. This is the case in massive courses like MOOCs or even when there are a lot of assignments during a course. One of the options to overcome this problem is to avoid open-response in favor of multiple-choice questions; but this way reduces significantly the communication between students and instructors that may involve handling different forms of data including computer programs, video, audio, and written texts. The alternative is that the students that authored the answers play also a role in the assessment. Peer assessment has been explored as an efficient procedure to deal with this problem, see for instance [6, 13, 14, 15, 16, 17, 7, 3, 10, 11, 9]. It has even acknowledged as an activity that enhances student learning in [18].

However, peer assessment has some flaws that should be addressed in order to be deployed more extensively. First, the quality of the feedback received by students should be improved [4, 8, 19]; in addition to a grade, students should obtain some annotations pointing to the weak and strong aspects of their answers. Second, peer assessment may increase considerably the burden on students.

In this paper we explore a method to tackle these two deficiencies when open-response are written documents. To improve the feedback we propose to use a set of labels or annotations that may be attached to answers with a level. These labels should cover the explanations that a student could

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Criteria	Levels										
The answer contains misspellings	○ many	○ some	○ few	● none							
Quality of the composition	○ bad	○ improvable	○ acceptable	● good							
Short term financial analysis	○ deficient	○ insufficient	○ sufficient	○ good	● excellent						
Long term financial analysis	○ deficient	○ insufficient	○ sufficient	● good	○ excellent						
Economical analysis	○ deficient	● insufficient	○ sufficient	○ good	○ excellent						
Global grade	○ 0	○ 1	○ 2	○ 3	○ 4	● 5	○ 6	○ 7	○ 8	○ 9	○ 10

Figure 1: Template used to annotate the assessment of the answers in the assignment of *Accounting Information*, see Section 4.1

obtain from a personalized assessment given by a professional instructor. We tested this proposal in 3 courses of our University from different fields: Laws and Economy. Instructors could easily express the possibilities of annotations in terms of labels with levels. On the other hand, the students understood effortlessly the assessment task with annotations.

The output of peer assessments is a dataset that must be filtered somehow to aggregate or reconcile the grades received by one answer from several students acting as graders without experience in this task. This is usually faced using Machine Learning methods. In the experiments reported at the end of the paper, we prove that models learned to aggregate grades can be used to ease the load of academic work of students.

The idea is to extend the assessment model to answers not involved in peer assessment in any way. For this purpose, we use a content-based approach similar to those used in Recommender Systems. Using a simple vectorial representation of the answers, we propose a matrix factorization method to learn how to grade. In fact, we have to learn to grade each of the aspects of the answer that need to be considered: the global grade, and the level of each of the labels or annotations for feedback. We present a multitask [2] method to learn simultaneously all the aspects to be assessed, and we show in the experiments that, in fact, there is an inductive transfer that improves the whole Machine Learning process.

The paper is organized as follows. First we explain the whole process as it is seen by students and instructors, and then we introduce the insight behind the approach presented here. Next a section is devoted to present the formal setting. Then we report the experiments conducted to evaluate the approach presented in the paper. We end with the conclusions of this research.

2 Overall Description of the Method

After submitting their answers for an open-response assignment, the students are required to grade a group of anonymized answers of other students. The assessments must be done using a template like that depicted in Figure 1, and following the rules detailed in a *rubric*. Notice that the template presents a set of graded annotations or labels that will form with the global grade the feedback for the students that authored the answers.

The rubric should include the *correct* answer, when this is not clear for all students; this was the case of the assignment of *Constitutional Law* used in the experiments reported in Section 4. But other times, for instance in the course titled *Spanish Economy*, the rubric indicates what elements would contain a good answer, and the grade is somehow a subjective opinion of the grader.

The labels and the rubric must be provided by the instructor considering that they serve to organize the process of assessment. They should aim to achieve uniform assessment criteria.

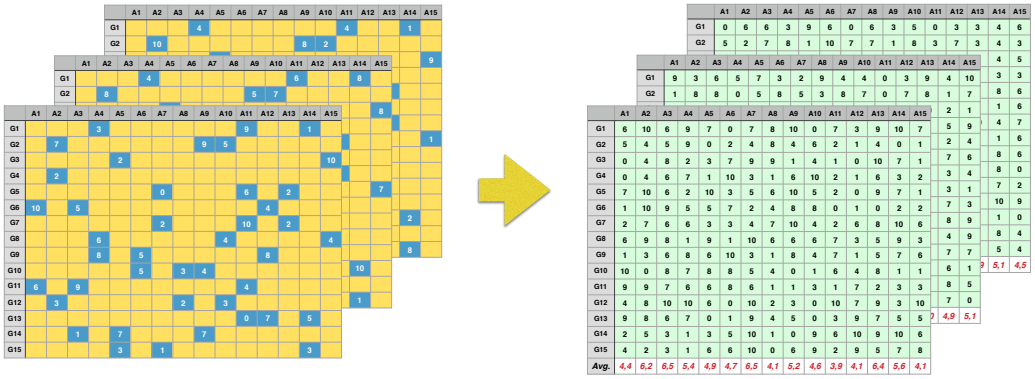

Figure 2: The process starts from a sparse assessment matrix and provides a full matrix after learning from the available data

The output of peer assessment is a 3-dimension matrix as that of Figure 2. In the figure we represented in rows the answers, in columns the graders, and in pages the labels to be graded. Typically, most of the components of this matrix are empty.

We assume that there is an unknown relation between the grades given to the labels of the answers and so if we find a pattern in grades of some of these labels, we hope to use them explicitly as an inductive transfer to learn how to make an assessment of all labels of all answers by all graders. In other words, we try to complete the assessment matrix with grades *consistent* with those we have available; see the right part of Figure 2.

The *consistency* of grades with the original assessment matrix is established in term of orderings. The aim is to have a ranking of answers as similar as possible to the partial rankings provided by graders. Thus, we are not going to use regression. The main reason is that graders are not professionals. Therefore, if a grader assigns 9 points to an answer $\boldsymbol{x}$ and 4 points to $\boldsymbol{y}$, we are only using the fact that $\boldsymbol{x}$ is *preferable to* $\boldsymbol{y}$. This is the *ordinal* point of view. If we were trying to learn how to predict exactly 9 points for $\boldsymbol{x}$ and 4 for $\boldsymbol{y}$ (the regression approach), then we would had adopt the *cardinal* point of view.

There are many reasons in favor of the ordinal approach, not only in assessment, but in general when we are interested in learning the preferences in contexts like information retrieval or marketing studies [1, 5, 12].

Once we have a complete estimation of the assessment matrix, for each label, we compute the average of all grades assigned to each answer in each label, including the global grade. These values will be the *grades* given by the model just learned from peer assessment data. However, those grades are just a tool to order the answers.

Sometimes, these rankings (one for each label) are enough to finish the assessment process. If this is not the case (as happens in the Spanish universities), we need to compute a grade. Then we transform ranking positions into grades following the same distribution as those granted by students acting as graders. In this sense we take into account the grades given by graders. But let us emphasize that this final step is just a translation from the language of percentiles to grades.

3 Formal Settings

Let $\mathcal{G}$ be a set of *graders*, and $\mathcal{A}$ a set of answers for an assignment. Graders are asked to assign a grade for a set of aspects of the assignment represented by labels in $\mathcal{L}$. After the assessment, we have an *assessment matrix*,

$$\boldsymbol{M}(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{a}). \tag{1}$$

The values of this matrix are grades given by a grader $\boldsymbol{g} \in \mathcal{G}$ for answer $\boldsymbol{a} \in \mathcal{A}$ with respect to a label $\boldsymbol{l} \in \mathcal{L}$. Typically, one of the labels stands for the *global grade*, but formally this is only another label to be assessed. The rest of labels will be understood as feedback given to the students who wrote the answers to explain the final grade.

Not all components of $\boldsymbol{M}$ are going to have values. In fact, frequently these type of matrices are quite sparse. The reason is that each grader $\boldsymbol{g}$ is asked to evaluate only a few answers $\mathcal{A}_{\boldsymbol{g}} \in \mathcal{A}$. As was mentioned above, the first step toward an assessment, is to fill the matrix according to the values available. For this purpose we start from a set of *preference judgments*, [1, 5, 12],

$$\mathcal{D} = \{(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{a}^b, \boldsymbol{a}^w) : \ \boldsymbol{M}(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{a}^b) > \boldsymbol{M}(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{a}^w)\}, \tag{2}$$

where $\boldsymbol{g} \in \mathcal{G}$, $\boldsymbol{l} \in \mathcal{L}$, and $\boldsymbol{a}^b, \boldsymbol{a}^w \in \mathcal{A}_{\boldsymbol{g}}$. The intended meaning is to record that for $\boldsymbol{g}$, for label $\boldsymbol{l}$, the answer $\boldsymbol{a}^b$ deserves a higher grade than the answer $\boldsymbol{a}^w$. In this way we overcome the actual grades but we retain the ordinal preferences of graders. Nevertheless, at the end of the process we will take into account the distribution of grades given for each label, just in case we need to transform the final ranking into absolute scores.

To handle answers, labels, and graders, we use a vectorial representation. Thus, for answers we use a *bag of words* approach to consider explicitly the contents of the answers in the assessment method. This requires that we first compute the *corpus* of all words used in all answers in $\mathcal{A}$. Then each answer can be codified by a binary vector indexed by the corpus: the components corresponding to a word that appears in the answer will have a value 1, while the rest will have a 0.

On the other hand, both graders and labels will be codified by binary vectors. The i^{th} element will be codified by a vector whose only nonzero value will be the i^{th} component. Then, to consider at the same time grader $\boldsymbol{g}$ and label $\boldsymbol{l}$ we use the *direct sum* (concatenation) of their vectorial representations, $(\boldsymbol{g} \oplus \boldsymbol{l})$.

All vectors involved in the assessment process will be projected (*embedded*) in a common Euclidean space $\mathbb{R}^k$,

$$\mathbb{R}^{|\mathcal{G}|+|\mathcal{L}|} \mapsto \mathbb{R}^k, \qquad (\boldsymbol{g} \oplus \boldsymbol{l}) \mapsto \boldsymbol{W}(\boldsymbol{g} \oplus \boldsymbol{l}), \tag{3}$$

$$\mathbb{R}^{|\operatorname{corpus}(\mathcal{A})|} \mapsto \mathbb{R}^k, \qquad \boldsymbol{a} \mapsto \boldsymbol{V}\boldsymbol{a}. \tag{4}$$

Notice that since the input of projections depend on the size of the Corpus and the number of labels and graders, $\mathbb{R}^k$ has an arbitrary dimension k. Typically we use a lower dimension than that of input spaces. The idea is to reduce the noise of the data.

In this context, we define a *full* assessment matrix $\widehat{\boldsymbol{M}}$ to estimate the grade for a label $\boldsymbol{l}$ given to an answer $\boldsymbol{a}$ according to grader $\boldsymbol{g}$, using the inner product of the projections in $\mathbb{R}^k$ as follows:

$$\widehat{\boldsymbol{M}}(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{a}) = \langle \boldsymbol{W}(\boldsymbol{g} \oplus \boldsymbol{l}), \boldsymbol{V}\boldsymbol{a} \rangle = (\boldsymbol{g} \oplus \boldsymbol{l})^T \boldsymbol{W}^T \boldsymbol{V}\boldsymbol{a}. \tag{5}$$

In this equation, the matrices $\boldsymbol{W}^T$ and $\boldsymbol{V}$ are *factors* of a matrix of weight for the products of the components of $(\boldsymbol{g} \oplus \boldsymbol{l})$and $\boldsymbol{a}$. For this reason, this apporach is called *matrix factorization*.

Finally, the grade for the aspect $\boldsymbol{l}$ of the answer $\boldsymbol{a}$ is defined by the average of grades given by all graders using the estimations of $\widehat{\boldsymbol{M}}$; see Figure 2. In symbols,

$$f(\boldsymbol{l}, \boldsymbol{a}) = \frac{1}{|\mathcal{G}|} \sum_{\boldsymbol{g} \in \mathcal{G}} \widehat{\boldsymbol{M}}(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{a}) = \langle \boldsymbol{W}(\bar{\boldsymbol{g}} \oplus \boldsymbol{l}), \boldsymbol{V}\boldsymbol{a} \rangle, \tag{6}$$

where $\bar{\boldsymbol{g}}$ stands for the *average grader*,

$$\bar{\boldsymbol{g}} = \frac{1}{|\mathcal{G}|} \sum_{\boldsymbol{g} \in \mathcal{G}} \boldsymbol{g}. \tag{7}$$

The coherence of the assessment matrix $\boldsymbol{M}$ and its estimation $\widehat{\boldsymbol{M}}$ is measured in terms of differences in the orderings of the answers. The aim is that the orderings induced by the estimated grades (of the average grader and each of the graders) are as similar as possible to the ordering given by each grader. Then we search for the best matrices $\boldsymbol{W}$ and $\boldsymbol{V}$. The formalization of our multitask approach is that both parameters, $\boldsymbol{W}$ and $\boldsymbol{V}$, are the same for all labels.

To measure the similarity of the orderings, we use a maximum margin approach. We pursue to reduce the number of swapped pairs in the orderings. The optimization problem considering all labels at the same time can be set to minimize the following loss function:

$$\operatorname{err}(\boldsymbol{W}, \boldsymbol{V}) = \sum_{(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{a}^b, \boldsymbol{a}^w) \in \mathcal{D}} \max\Big\{0, 1 - \langle \boldsymbol{W}((\bar{\boldsymbol{g}} + \boldsymbol{g}) \oplus \boldsymbol{l}), \boldsymbol{V}\boldsymbol{a}^b \rangle + \langle \boldsymbol{W}((\bar{\boldsymbol{g}} + \boldsymbol{g}) \oplus \boldsymbol{l}), \boldsymbol{V}\boldsymbol{a}^w \rangle\Big\}. \tag{8}$$

To solve this optimization problem we use a Stochastic Gradient Descent (*SGD*), that in each iteration updates the parameters of the model, Θ (in this case the matrices $\boldsymbol{W}$ y $\boldsymbol{V}$) as follows:

$$\Theta \leftarrow \Theta - \gamma \left(\frac{\partial \operatorname{err}(\Theta)}{\partial \Theta} + \nu \cdot \frac{\partial \|\Theta\|_F^2}{\partial \Theta} \right), \tag{9}$$

where $\| \cdot \|_F^2$ is the Frobenius norm included for *regularization*, γ is the *learning rate* and ν is the *regularization factor*. As usual γ decreases its value in each iteration; in the experiments reported at the end of this paper, to determine the value of γ in the i^{th} iteration, we have used the expression

$$\gamma = \frac{1}{1 + \gamma_s \cdot i}. \tag{10}$$

The derivatives used in the SGD, when the maximum in (8) is greater than zero, are given by

$$\frac{\partial \operatorname{err}(\Theta)}{\partial \boldsymbol{W}} = \boldsymbol{V}(\boldsymbol{a}^w - \boldsymbol{a}^b)((\bar{\boldsymbol{g}} + \boldsymbol{g}) \oplus \boldsymbol{l})^{\mathrm{T}} \tag{11}$$

$$\frac{\partial \operatorname{err}(\Theta)}{\partial \boldsymbol{V}} = \boldsymbol{W}((\bar{\boldsymbol{g}} + \boldsymbol{g}) \oplus \boldsymbol{l})(\boldsymbol{a}^w - \boldsymbol{a}^b)^{\mathrm{T}} \tag{12}$$

3.1 Binary Relevance (BR)

The straightforward baseline for the multitask approach is to learn one model for each label. In our case, to learn matrices $\boldsymbol{W_l}$ and $\boldsymbol{V_l}$ for each label $\boldsymbol{l}$. For this purpose we need to focus only on those grades involving one label $\boldsymbol{l}$,

$$\mathcal{D}_{\boldsymbol{l}} = \{(\boldsymbol{g}, \boldsymbol{a}^b, \boldsymbol{a}^w) : \ (\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{a}^b, \boldsymbol{a}^w) \in \mathcal{D}\}. \tag{13}$$

Using a methodology parallel to that presented above for the multitask approach, we estimate the grades for an answer $\boldsymbol{a}$ and a label $\boldsymbol{l}$ using a particular function

$$f_{\boldsymbol{l}}(\boldsymbol{a}) = \langle \boldsymbol{W_l}\bar{\boldsymbol{g}}, \boldsymbol{V_l}\boldsymbol{a} \rangle. \tag{14}$$

In the following we will refer to this simple approach as *Binary Relevance* (BR) using a terminology borrowed from *multilabel* classification.

From a geometrical point of view, the grades given both by f (6) and $f_{\boldsymbol{l}}$ (14) are inner products in $\mathbb{R}^k$. For a label $\boldsymbol{l}$, the assessment of an answer $\boldsymbol{a}$ is proportional to the distance from $\boldsymbol{V}\boldsymbol{a}$ or $\boldsymbol{V_l}\boldsymbol{a}$ to a hyperplane in $\mathbb{R}^k$. The hyperplane is defined as the perpendicular to the projection of a vector that involves only the label $\boldsymbol{l}$: $\boldsymbol{W}(\bar{\boldsymbol{g}} \oplus \boldsymbol{l})$ in multitask approach, and $\boldsymbol{W_l}\bar{\boldsymbol{g}}$ in *Binary Relevance*.

Thus, the work entrusted to matrices $\boldsymbol{V}$ and $\boldsymbol{V_l}$ is to place answers $\boldsymbol{a}$ in $\mathbb{R}^k$ in such a way that the distances to some hyperplanes are coherent with the ordering of graders. Figure 3 shows an example where $\mathbb{R}^k$ has $k = 2$ and the hyperplanes are red lines.

In this context, multitask approach places the representations of answers (blue points in the figure) in the same position for all labels. But Binary Relevance starts from scratch for each label and so the clues given by one label can not be used in any way to place the projections of answers for another label.

3.2 Transforming the Ranking into a Grade

After learning the matrices $\boldsymbol{W}$ and $\boldsymbol{V}$, using the function f (6), for each label, we have a ranking of answers from best to worse. If we need to transform this ranking into grades, our proposal is to do that trying to reproduce the same distribution of grades that we collected from students. Notice that this is only a translation that has no effect in the ranking of answers learned in the multitask described above.

Of course, for BR we may follow an analogous process to obtain grades for each answer in each label.

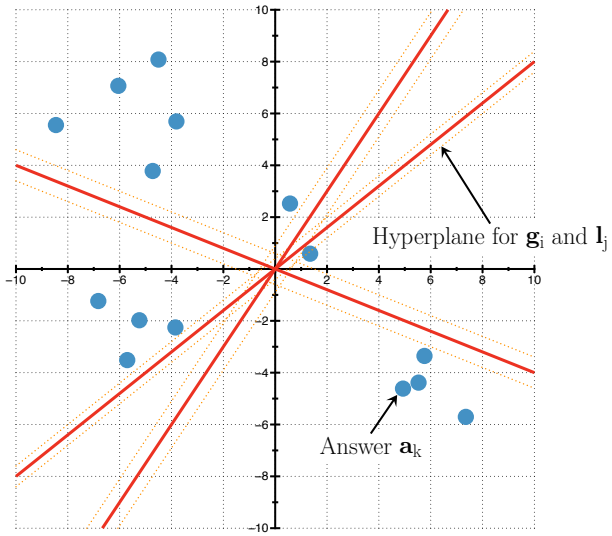

Figure 3: Geometrical interpretation of the multitask approach. The blue points represent answers, while the red lines are the hyperplanes defined by the assessment labels. The location of answers is learned to cope with all labels as the same time aiming to take advantage of an inductive transfer between them

4 Experimental Results

In this section we report a number of experiments performed to test the goodness of the method presented in this paper. First we introduce the datasets used, then the evaluation method, and finally the scores obtained.

4.1 Datasets

The datasets used in the experiments were gathered from 3 courses of different fields in the *Universidad de Oviedo*: Accounting Information, Constitution Law, and Spanish Economy.

To collect the data we used a Moodle (moodle.org) installation in one of our servers. This platform has a tool called *workshop* that provides the infrastructure required for peer assessment. The final grade is computed in this tool by averaging the grade received by each answer, thus we replaced this step by our method.

The assessment was double-blind guaranteeing also that no student graded her or his own answer. Each student received 10 answers to grade.

Table 1 shows the basic characteristics of the datasets. Note that around 90% of the components of the assessment matrices $\boldsymbol{M}$ (1) are empty.

Table 1: Characteristics of the datasets

	Account Information	Constitutional Law	Spanish Economy
Number of answers	119	66	111
Number of graders	112	66	108
Number of assessments	1120	660	1065
Empty (%)	92.09	84.85	91.36
Average number of grades per answer	9.41 ± 0.71	10 ± 0	9.59 ± 0.67
Average number of grades per grader	10 ± 0	10 ± 0	9.86 ± 0.99

Table 2: Detailed description of grades and labels for each dataset

Accounting Information

#PJ	Discrepancies (%)	Labels
1603	3.74	(1) The answer contains misspellings
3068	5.05	(2) Quality of the composition
3043	4.24	(3) Short term financial analysis
3187	4.61	(4) Long term financial analysis
3455	4.08	(5) Economical analysis
4233	5.20	Global grade

Constitutional Law

#PJ	Disrepancies. (%)	Labels
570	2.98	(1) The answer contains misspellings
1273	6.44	(2) Quality of the composition
1172	4.95	(3) Line of arguments
378	2.91	(4) Quotes the relevant papers
171	0.00	(5) Does not know what is a motion of censure
218	2.29	(6) Does not know what is a motion of non-confidence
369	1.90	(7) Does not know the duties of the King
184	0.00	(8) Does not know how the President is appointed
112	2.68	(9) Does not know the duties of the President
2158	9.73	Global grade

Spanish Economy

#PJ	Disrepancies. (%)	Labels
2318	5.95	(1) Ability to understand and expose the core economic processes of each of the stages of evolution of the Spanish economy
2331	5.19	(2) Ability to distinguish the phases of convergence and divergence of the Spanish economy on the European economy
2329	5.84	(3) Ability to show the overall balance of the evolution of the Spanish economy with its main achievements and limitations
2544	6.29	(4) Ability to reasonably explain the salient features, events and consequences of the recent economic crisis and the dilemmas posed to economic policies
2735	5.45	(5) Quotes the relevant references and incorporates well-reasoned personal judgment
2648	6.50	(6) The argument are well organized and clear. The answer shows synthesis capacity and use the right economic terms
3736	8.00	Global grade

On the other hand, Table 2 shows additional characteristics of the datasets. The first column (*#PJ*) reports the number of *Preference Judgments*; that is, the size of the corresponding dataset $\mathcal{D}_l$ (13). Notice that for each label we record only those pairs of answers with different grade given by a grader; the pairs with the same grade do not give rise to any element in $\mathcal{D}_l$. This is the reason why the number of preference judgments may be different for different labels. Recall that multitask approach deals with the join of $\mathcal{D}_l$ for all l in a single $\mathcal{D}$ (2).

The second column (*Discrepancies*) collects the percentage of contradictory preferences with the majority of opinions. For instance, if for a label l 3 graders think that answer x is better than y and other 2 graders think the opposite, we count 2 discrepancies. Thus, the percentage of discrepancies is a lower error bound for any classification function.

Finally, the last column of Table 2 details the set of labels used in the assignments whose data was used in the experiments. The original were written in Spanish, so here we give a translation. The number in parentheses is the same used in Figure 4. This picture represents the distribution of grades given by graders for all labels including the global grade in the rightmost graphic of each row.

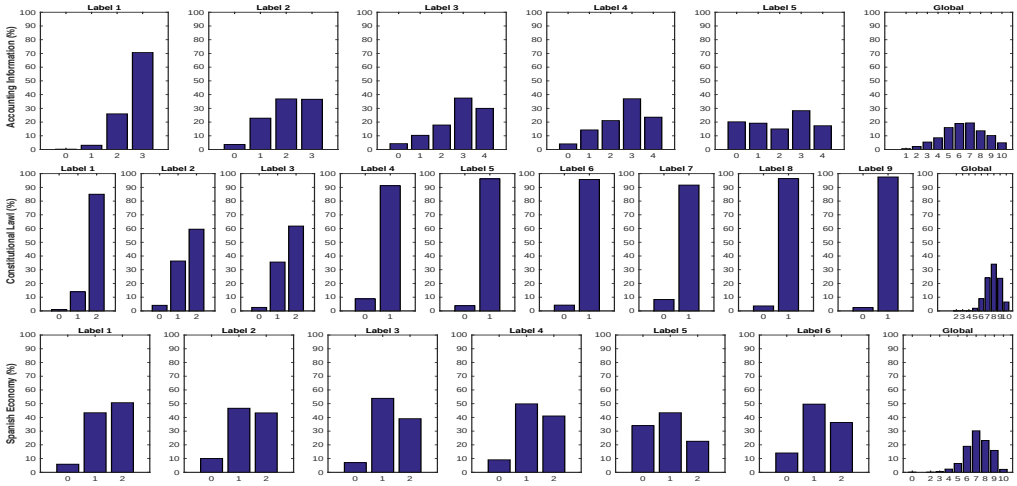

Figure 4: Distributions given by graders for all labels

4.2 Evaluation Method

To evaluate the performance of the multitask approach presented in this paper, we made some train/test experiments with the datasets described above. We compared the performance of multitask versus BR (Section 3.1). To split the datasets we first separated a set of students and made a train set with only the preference judgments involving this subset of students, either as graders or as authors of the answers. The rest of available preference judgments were then considered as test set. The size of the set of students selected was 25, 50, 75 and 100, except in the case of the dataset from Constitutional Law since we only had 66 students and therefore we only considered training sets of students of size 25 and 50.

The performance measure was a simple 0/1 classification error in test. Errors are those ordered pairs of answers in test sets that were not ordered in the same way by the function f (6) learned by the multitask approach. Table 3 shows the percentage of errors computed averaging 10 repetitions.

During training, the SGD algorithm uses some parameters that must be set in order to ensure the best performance. For this purpose, we made a grid search using only training sets to find the most promising combination. We made a cross-validation experiment with 2 folds and 5 repetitions using all possible combinations of values of k, ν and γ_s (10):

$$\begin{aligned} k &\in \{2, 10, 20, 50, 100\}, \\ \nu &\in \{10^e : e = -4, \ldots + 2\}, \\ \gamma_s &\in \{10^e : e = -4, \ldots, -1\}. \end{aligned} \tag{15}$$

Then we selected the best combination to perform the corresponding train/test experiment.

4.3 Results

Table 3 report the scores achieved by the multitask approach and the BR approaches. The last row of each table shows the weighted average of the scores of all labels for each train/test; the weights are the number of test elements.

In boldface we highlight the best (weighted average) scores. We observe that in most cases the multitask outperforms the BR approach: 8 out of 10 times multitask is better.

As expected, the error decreases as the size of the training set increases. The scores attained in the biggest sets are slightly greater than those reported in papers like [14, 10, 11, 9]. The main reason is than in those cases the error reported were resubstitutions; the comparison was established comparing the discrepancies between profesional instructors and the model learned.

In this paper we present a collection of train/test experiments that, on the one hand, provide a support to launch assessment tools where only a part of the students will be required to grade. On the other

Table 3: Percentage of errors

Accounting Information

	Multitask				BR			
	25	50	75	100	25	50	75	100
L1	43.29	45.79	48.10	44.26	49.47	47.74	46.08	44.00
L2	36.86	31.24	36.14	28.75	45.22	32.41	32.78	27.74
L3	48.24	30.96	30.44	23.72	48.26	33.01	27.88	23.89
L4	41.91	35.86	32.99	29.42	47.24	36.27	32.08	30.60
L5	40.50	30.27	31.57	25.53	44.57	33.57	29.90	26.34
Global	40.22	29.60	28.21	25.30	45.85	30.09	26.67	25.30
weighted	**41.58**	**32.68**	32.99	**27.96**	46.45	34.17	**31.03**	28.15

Constitucional Law

	Multitask		BR	
	25	50	25	50
L1	43.24	36.82	39.71	38.32
L2	34.07	33.90	35.73	37.45
L3	28.97	30.41	30.76	34.33
L4	35.80	23.63	46.56	27.81
L5	43.01	36.81	51.70	36.39
L6	31.39	33.08	43.56	27.05
L7	42.28	18.42	40.73	23.51
L8	36.47	26.43	42.88	30.86
L9	26.87	31.43	44.34	30.86
Global	34.95	36.40	36.92	35.62
weighted	**34.88**	**33.04**	37.46	34.68

Spanish Economy

	Multitask				BR			
	25	50	75	100	25	50	75	100
L1	54.94	40.65	39.71	32.13	49.57	47.98	43.49	38.03
L2	52.46	40.72	36.75	37.07	43.24	45.41	40.85	39.66
L3	51.54	40.72	36.69	30.78	47.08	44.65	41.86	37.51
L4	50.44	42.61	40.60	31.51	48.42	46.17	43.98	36.31
L5	42.62	36.23	35.92	31.25	45.09	41.04	37.20	35.81
L6	49.14	43.04	40.53	40.38	49.26	48.33	43.01	39.58
Global	49.06	40.63	38.40	33.61	48.23	46.41	38.92	37.96
weighted	49.78	**40.62**	**38.38**	**33.79**	**47.34**	45.72	41.11	37.82

hand, the experiments reported here back the hypothesis that assessment can be smoothly learned like other learning tasks.

5 Conclusions

In this paper we address two important problems to increase the quality of peer assessment of written open-responses: provide useful feedback to students, and relieve their workload. The proposal requieres the graders to assess some annotations or labels about the answer that they are assessing. The global grade is another label in this context. We have presented a method that uses a multitask approach to search for grading patterns in all labels at the same time.

Multitask leverages the accuracy of a baseline that focus successively on each label separately. Thus, the assessments provided by students can be aggregated in a list of graded labels that informs students of their global grade, as well as of a number of reasons to explain weak and strong points of their answers.

On the other hand, models learned with the multitask approach can be extended to answers not involved at all in peer assessment. The consequence is that a part of students can be relieved of the assessment task reducing in this way the burden on students in this kind of processes.

The goodness of the approach presented in this paper was checked on 3 datasets collected in courses of our University (Accounting Information, Constitutional Law, and Spanish Economy) yielding quite successful accuracy scores. Therefore, we would like to underscore that this research proved that it is feasible to deploy sophisticated assessment methods in fields far from Computer Science. Both instructors and students found the experience satisfactory and they did not find any difficulty in moving from traditional assignments to our proposal.

Acknowledgments

The research reported in this paper has been supported in part under a MINECO/FEDER grant TIN2015-65069-C2-2-R from the Ministerio de Economía y Competitividad and partially funded by FEDER (European Regional Development Fund). We also would like to acknowledge students and instructors who collaborated with us in the assignments of our University, Universidad de Oviedo, in Spanish Economy (Juan Vázquez), Constitutional Law (Francisco Bastida) and Accounting Information (Mónica Álvarez Pérez).

References

[1] Antonio Bahamonde, Gustavo F. Bayón, Jorge Díez, José Ramón Quevedo, Oscar Luaces, Juan José del Coz, Jaime Alonso, and Félix Goyache. Feature subset selection for learning preferences: A case study. In Russ Greiner and Dale Schuurmans, editors, *Proceedings of the International Conference on Machine Learning (ICML '04)*, pages 49–56, Banff, Alberta (Canada), July 2004.

[2] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[3] Jorge Díez, Oscar Luaces, Amparo Alonso-Betanzos, Alicia Troncoso, and Antonio Bahamonde. Peer Assessment in MOOCs Using Preference Learning via Matrix Factorization. In *NIPS Workshop on Data Driven Education*, 2013.

[4] Sarah Gielen, Elien Peeters, Filip Dochy, Patrick Onghena, and Katrien Struyven. Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, 20(4):304 – 315, 2010. Unravelling Peer Assessment.

[5] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2002.

[6] Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. Peer and self assessment in massive online classes. In Hasso Plattner, Christoph Meinel, and Larry Leifer, editors, *Design Thinking Research*, Understanding Innovation, pages 131–168. Springer International Publishing, 2015.

[7] Igor Labutov and Christoph Studer. JAG: Joint Assessment and Grading. In *Machine Learning for Digital Education and Assessment Systems, ICML 2016 Worshop*, 2016.

[8] Ngar-Fun Liu and David Carless. Peer feedback: the learning element of peer assessment. *Teaching in Higher education*, 11(3):279–290, 2006.

[9] Oscar Luaces, Jorge Díez, Amparo Alonso, Alicia Troncoso, and Antonio Bahamonde. Including content-based methods in peer-assessment of open-response questions. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 273–279. IEEE, 2015.

[10] Oscar Luaces, Jorge Díez, Amparo Alonso-Betanzos, Alicia Troncoso, and Antonio Bahamonde. A factorization approach to evaluate open-response assignments in MOOCs using preference learning on peer assessments. *Knowledge-Based Systems*, 85:322 – 328, 2015.

[11] Oscar Luaces, Jorge Díez, Amparo Alonso-Betanzos, Alicia Troncoso, and Antonio Bahamonde. Content-based methods in peer assessment of open-response questions to grade students as authors and as graders. *Knowledge-Based Systems*, 2016.

[12] Oscar Luaces, Jorge Díez, Thorsten Joachims, and Antonio Bahamonde. Mapping preferences into euclidean space. *Expert Systems with Applications*, 42(22):8588 – 8596, 2015.

[13] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. Tuned models of peer assessment in MOOCs. In *Proceedings of the 6^{th} International Conference on Educational Data Mining (EDM'13)*, pages 153–160. International Educational Data Mining Society, 2013.

[14] Karthik Raman and Thorsten Joachims. Methods for ordinal peer grading. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2014.

[15] Karthik Raman and Thorsten Joachims. Bayesian ordinal peer grading. In *Proceedings of the Second (2015) ACM Conference on Learning Scale*, pages 149–156, New York, NY, USA, 2015. ACM.

[16] Philip M Sadler and Eddie Good. The impact of self-and peer-grading on student learning. *Educational Assessment*, 11(1):1–31, 2006.

[17] Nihar B Shah, Joseph K Bradley, Abhay Parekh, Martin Wainwright, and Kannan Ramchandran. A case for ordinal peer-evaluation in MOOCs. In *NIPS Workshop on Data Driven Education*, 2013.

[18] Dennis L Sun, Naftali Harris, Guenther Walther, and Michael Baiocchi. Peer assessment enhances student learning: The results of a matched randomized crossover experiment in a college statistics class. *PloS one*, 10(12):e0143177, 2015.

[19] Sheng-Chau Tseng and Chin-Chung Tsai. On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education*, 49(4):1161 – 1174, 2007.