Detection of Inconsistencies in Student Evaluations

Štefan Pero and Tomáš Horváth
Institute of Computer Science, Faculty of Science, University of Pavol Jozef Šafárik, Košice, Slovakia

Keywords: Grading, Student Assessment, Inconsistent Evaluation, Textual Evaluation, Personalization.

Abstract: Evaluation of the solutions for the tasks or projects solved by students is a complex process driven mainly by the subjective evaluation criteria of a given teacher. Each teacher is somehow biased meaning how strict she is in assessing grades to solutions. Besides the teacher's bias there are also some other factors contributing to grading, for example, teachers can make mistakes, the grading scale is too rough-grained or too fine-grained, etc. Grades are often provided together with teacher's textual evaluations which are considered to be more expressive as a single number. Such textual evaluations, however, should be consistent with grades, meaning that if two solutions have very similar textual evaluations their grades should be also very similar. Though, some inconsistencies between textual evaluations and grades provided by the teacher used to arise, especially, when a teacher has to assess a large number of solutions, or if more than one teacher is involved in the evaluation process. We propose a simple approach for detection of inconsistencies between textual evaluations and grades in this paper. Experiments are provided on two real-world datasets collected from the teaching process at our university.

1 INTRODUCTION

The way how a teacher grades students' tasks or projects is a complex process depending on the teacher's *subjective evaluation* criteria. However, teachers are usually not provided with standards for grading, only some district or school policies offer some guidance for teachers (Banta et al., 2009; Walvoord and Anderson, 2009). Moreover, in many cases, an evaluation has to be done on a fine-grained scale (e.g. from 0 to 100) facilitating a teacher to grade two very similar or even equal solutions slightly differently. On the other hand, a roughly grained grading scale (e.g. from 1 to 5) often forces the teacher to under-evaluate or over-evaluate student works because the grading scale does not allow her to give a rating in between some certain two values. What difference does it really make if the grade is 2- or 3+ instead of 2 or 3, respectively (Carell and West, 2010)? Evaluation is a highly inconsistent process. Teachers have various types of evaluation and assessment criteria they give different values to and weight them differently (Suskie, 2009; Rockoff and Speroni, 2010).

Grading also tends to reduce students interest in the learning itself. Students tend to lose interest in whatever they have to do, instead they are rather focused to get a grade (Kohn, 1999). Results of some research demonstrated that "grade orientation" and "learning orientation" are inversely related, and that grading also tends to reduce students preferences for challenging tasks what affects the quality of students thinking (Beck et al., 1991; Milton et al., 1986; Milton, 2009).

In many cases, grading are provided together with a *textual evaluation*, a kind of a review, i.e. teacher's comments or complains to solutions. We think that a textual comment represents a more precise and expressive evaluation as a single number (a grade) because the teacher has the ability to express her attitude to the given solution in a more detailed view. However, textual evaluation is basically considered just as a feedback for the students and as a justification for the grading. In official reports, the final mark is used, though.

There is one important issue which should be taken into account here, namely, that textual evaluations should be consistent with the grades. It means that if the teacher provides very similar reviews for two solutions, the corresponding grades should be also very similar. This is especially important when there are more teachers evaluating students' solutions for the same lecture/topic since each teacher is somehow *biased*, i.e. some of them evaluate very strictly,

Pero Š. and Horváth T..
Detection of Inconsistencies in Student Evaluations.
DOI: 10.5220/0004385602460249
In *Proceedings of the 5th International Conference on Computer Supported Education* (CSEDU-2013), pages 246-249
ISBN: 978-989-8565-53-2

some of them are more friendly, etc. The problem of biases is well-known in recommender systems, especially in rating prediction (Koren et al., 2009). Real data shows[1] that there is a variance in rating biases w.r.t. the opinions expressed in product reviews even in case of a single user. i.e. a user overrates some items relative to the opinions/sentiments provided in her textual reviews for these items, while in the case of other items it is the opposite.

We were motivated in this work by a *real example* from one course at our university. The "Programming, Algorithms and Complexity" lecture is provided by one lecturer, however, the tutorials are realized by five assistants, each of them leading one group[2] of about 15-20 students. It is important that there are no two solutions with the same or very similar textual evaluations but quite different grades.

This work focuses on finding such inconsistencies in teachers' gradings according to their textual evaluations. The resulting set of inconsistent grade assessments can then be used to unify the evaluation process within a large course taught by more teachers, or in cases when one teacher has to evaluate a large number of solutions. The contributions of this work are the following:

- We introduce a formal model for the problem of inconsistency detection in teachers' evaluations. To the best of our knowledge, this work is the first one devoted to this problem.
- We propose a simple approach to detect inconsistent evaluations utilizing TF-IDF, a well-known techniques from information retrieval. We illustrate the complete process of inconsistency detection on two real-world datasets, both collected from our colleagues at our university.

2 INCONSISTENT EVALUATIONS

In our formal model, we define a *task* as a triple $\mathbf{t} = (\sigma, \pi, \varsigma)$, where σ, π and ς refer to student, problem to be solved, and, the the provided solution for the given problem by the given student, respectively.

Teacher's *evaluation* is represented as a quadruple $\mathbf{e} = (\tau, \mathbf{t}, \theta, \gamma)$ where τ refers to teacher, $\mathbf{t}$ refers to task as defined above, and, θ, γ refer to the textual evaluation and grade, respectively, assigned by the given teacher to the given task. The textual evaluation $\theta = (w_1, w_2, \ldots, w_k)$ is basically a sequence of words (terms) in the same order as they appear in the text, while $\gamma \in \mathbb{Q}$ is some rational[3] number. From now on, we will call textual evaluations as reviews.

Information collected about the concrete entities of σ, π, ς or τ are not necessary for inconsistency detection (see the proposed approach below), however, it is good to have these information captured in our model to be able to derive some additional facts, such as, which teachers are the highest inconsistency among, which problems to be solved cause high inconsistencies between evaluators, etc.

The set of evaluations, the input for our inconsistency detection approach introduced here, is denoted as $\mathcal{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$.

The set $\mathcal{I}$ of inconsistent evaluations contains those pairs of evaluations for which the textual parts (the reviews) are similar but the grades differ.

$$\mathcal{I} = \{(\mathbf{e}_i, \mathbf{e}_j) \mid sim(\theta_i, \theta_j) \geq \delta,\ dif(\gamma_i, \gamma_j) \geq \varepsilon\} \quad (1)$$

where $sim(\theta_i, \theta_j)$ is a similarity of the reviews $\theta_i \in \mathbf{e}_i$ and $\theta_j \in \mathbf{e}_j$, and $dif(\gamma_i, \gamma_j)$ denotes the difference in grades $\gamma_i \in \mathbf{e}_i$ and $\gamma_j \in \mathbf{e}_j$.

2.1 Similarity of Textual Evaluations

In order to compute the similarity of reviews, first we need to represent these reviews in an appropriate way. For this purpose we use the *TF-IDF* function (Spärck Jones, 1972; Robertson, 2004; Wu et al., 2008) that stands for *term frequency–inverse document frequency*. TF-IDF is often used in information retrieval and text mining (Ramos, 2003) for measuring how important a word is related to a review in a collection of reviews. TF-IDF is the product of two statistics, term frequency and inverse document frequency.

In the case of term frequency $TF(w, \theta)$ is simply defined as the proportion of the raw frequency of the term w in the review θ and the maximal frequency of any term w' in the review θ:

$$\mathrm{TF}(w, \theta) = \frac{freq(w, \theta)}{max\{freq(w', \theta) \mid w' \in \theta\}} \quad (2)$$

The inverse document frequency is a measure of whether the term w is common across all the reviews and is defined as the ratio of the number of all reviews to the number of reviews containing the term w:

$$\mathrm{IDF}(w, \mathcal{E}^{\theta}) = \frac{|\mathcal{E}^{\theta}|}{|\{\theta' \in \mathcal{E}^{\theta} \mid w \in \theta'\}|} \quad (3)$$

[1] E.g. Amazon Product Review Data (Jindal and Liu., 2008), `http://liu.cs.uic.edu/download/data/`

[2] Student groups are organized according to the capacities of computer rooms and the programming skills of students.

[3] Grades are often representing percentages or some ratios of acquired points related to the maximum number of points. Thus, it is natural to consider rational numbers.

where $\mathcal{E}^{\theta} = \{\theta_i \mid \theta_i \in \mathbf{e}_i, \mathbf{e}_i \in \mathcal{E}\}$ is the set of (textual) reviews appearing in the evaluations in $\mathcal{E}$.

Each review $\theta \in \mathcal{E}^{\theta}$ is represented as a m-dimensional vector of TF-IDF scores $\theta^{\text{TF-IDF}} = (\text{TF-IDF}(w'_1, \theta, \mathcal{E}^{\theta}), \ldots \text{TF-IDF}(w'_m, \theta, \mathcal{E}^{\theta}))$, where $w'_1, \ldots, w'_m \in \mathcal{W} = \{w' \in \theta \mid \theta \in \mathcal{E}^{\theta}\}$ are all the terms appearing in all the reviews, and the TF-IDF score is calculated as:

$$\text{TF-IDF}(w, \theta, \mathcal{E}^{\theta}) = \text{TF}(w, \theta) \,.\, \text{IDF}(w, \mathcal{E}^{\theta}) \qquad (4)$$

Since, each review θ is represented as a m-dimensional vector $\theta^{\text{TF-IDF}}$, we define the similarity of two reviews θ_i, θ_j as their *cosine similarity* (Tan et al., 2005), a well-known vector similarity measure

$$sim(\theta_i, \theta_j) = \frac{\sum_{l=1}^{m} \theta_{i_l}^{\text{TF-IDF}} \theta_{j_l}^{\text{TF-IDF}}}{\sqrt{\sum_{l=1}^{m} \theta_{i_l}^{\text{TF-IDF}^2}} \sqrt{\sum_{l=1}^{m} \theta_{j_l}^{\text{TF-IDF}^2}}} \qquad (5)$$

2.2 Difference of Grades

When computing the difference between two grades γ_i and γ_j, we have to take into account also the size of the grading scale which is an interval $[\gamma_{min}, \gamma_{max}] \subset \mathbb{Q}$ of grades with $\gamma_{min}, \gamma_{max}$ being the minimal and maximal possible grades, respectively. The difference of two grades γ_i and $\gamma_j \in [\gamma_{min}, \gamma_{max}]$ is computed as

$$dif(\gamma_i, \gamma_j) = \frac{|\gamma_i - \gamma_j|}{\gamma_{max} - \gamma_{min}} \qquad (6)$$

2.3 Inconsistency Detection

Before introducing the algorithm, similarly to $\mathcal{E}^{\theta}$, we define the set $\mathcal{E}^{\gamma} = \{\gamma_i \mid \gamma_i \in \mathbf{e}_i, \mathbf{e}_i \in \mathcal{E}\}$ of grades appearing in the evaluations in $\mathcal{E}$. The above defined set $\mathcal{I}$ of inconsistent evaluations can be computed by the following algorithm:

Algorithm 1: Inconsistency detection.

Input: $\mathcal{E}, \mathcal{E}^{\theta}, \mathcal{E}^{\gamma}, \varepsilon, \delta$
Output: $\mathcal{I}$

```
for i = 1 to n − 1 do
  for j = i + 1 to n do
    sim_reviews ← sim(θ_i, θ_j), θ_i, θ_j ∈ E^θ
    dif_grades ← dif(γ_i, γ_j), γ_i, γ_j ∈ E^γ
    if sim_reviews ≥ ε and dif_grades ≥ δ then
      I ← I ∪ {(e_i, e_j)}
    end if
  end for
end for
return I
```

3 EXPERIMENTS

We have used two real-world datasets, the first labeled "PAC"[4] and the second labeled "PALMA"[5]. Both datasets contain the following information about evaluations: *studentID, taskID, teacherID, grade, review*. For our experiments, however, we used only the following tuples: $(ID, grade, review)$, where ID is a unique identifier created from *studentID*, *taskID* and *teacherID*. The main characteristics of the datasets are described in table 1.

Table 1: Characteristics of the datasets used.

Dataset	#Students	#Tasks	#Instances
PAC	82	18	174
PALMA	141	154	1501

First, we computed similarity of reviews using the TF-IDF measure as defined in the equation (4). This technique has also filtered "unusable" and "rarely used" words in reviews which we could omit from the consideration making the computation faster. In the next step we computed the set of inconsistent pairs of evaluations according to the algorithm 1.

The results shown in figures 1 and 2 refer to the number of inconsistent evaluations found in the data for different ε and δ. Since the PAC dataset is smaller, naturally we have found less inconsistencies than in the case of the PALMA dataset in which there are 8 pairs of evaluations ($|\mathcal{I}| = 8$) where the evaluated tasks are the same and the textual reviews for these tasks are equal ($\varepsilon = 1$) but their numerical evaluations (grades) differ with more than 30% ($\delta = 0.3$).

ε	δ	$\lvert\mathcal{I}\rvert$
1	0.05	0
1	0.1	0
1	0.3	0
0.95	0.05	4
0.95	0.1	4
0.95	0.3	1

ε	δ	$\lvert\mathcal{I}\rvert$
0.90	0.05	4
0.90	0.1	3
0.90	0.3	1
0.85	0.05	18
0.85	0.1	16
0.85	0.3	6

Figure 1: Results for inconsistency detection for various ε and δ in the PAC dataset.

The choice of the concrete values for δ and ε for our algorithm depends on the individual requirements

[4]Collected from the "Programming Algorithms Complexity" course at the Institute of Computer Science at Pavol Jozef Šafárik University during the years 2010–2012.

[5]Collected from the "PALMA junior" programming competition organized by the Institute of Computer Science at Pavol Jozef Šafárik University during the years 2005–2012.

ε	δ	$\|I\|$
1	0.05	8
1	0.1	8
1	0.3	8
0.95	0.05	13
0.95	0.1	11
0.95	0.3	9

ε	δ	$\|I\|$
0.90	0.05	13
0.90	0.1	10
0.90	0.3	8
0.85	0.05	14
0.85	0.1	12
0.85	0.3	7

Figure 2: Results for inconsistency detection for various ε and δ in the PALMA dataset.

of lecturers. However, providing results for different combinations of values of δ and ε (as in the figures 1 and 2) allows the teachers to gain better insight to the evaluation process of their lectures.

4 CONCLUSIONS

Teachers should evaluate students' solutions consistently, however, this is not always the case. We proposed a simple and easy to implement solution for detecting inconsistencies in the evaluation process when the textual review of two solutions provided for the same task are very similar but the numerical grades differ. Since, to the best of our knowledge, our work is the first dealing with this issue, we also introduced a formal model of the inconsistency detection problem. Experiments on two real-world datasets show that even in a small scale we can found inconsistent evaluations.

We provided our findings to the colleagues who provided us with the datasets as well as to some of our other colleagues at our university. Positive feedbacks from these teachers show that the introduced approach for evaluation inconsistency detection is helpful in the teaching process and worth further investigation.

Our further research will focus on the relationship between assessment methods and the learning outcomes of students, as well as the investigation of utilizing different feature extraction methods (Petz et al., 2012; Holzinger et al., 2012) in our approach.

ACKNOWLEDGEMENTS

This work was supported by the grants VEGA 1/0832/12 and VVGS-PF-2012-22 at the Pavol Jozef Šafárik University in Košice, Slovakia. We would like to thank to our colleagues František Galčík for providing us the PAC dataset, Ľubomír Šnajder and Ján Guniš for providing us the PALMA dataset.

REFERENCES

Banta, T. W., Jones, E. A., and Black, K. E. (2009). *Designing Effective Assessment: Principles and Profiles of Good Practice*. John Wiley and Sons, 2nd edition.

Beck, H. P., Rorrer-Woody, S., and Pierce, L. G. (1991). The relations of learning and grade orientations to academic performance. In *Teaching of Psychology 18*, pages 35–37.

Carell, S. E. and West, J. E. (2010). Random assignment of students to professors. *Journal of Political Economy*.

Holzinger, A., Yildirim, P., Geier, M., and Simonic, K.-M. (2012). *Quality-based knowledge discovery from medical text on the Web Example of computational methods in Web intelligence*. Springer.

Jindal, N. and Liu., B. (2008). Opinion spam and analysis. In *Proceedings of First ACM International Conference on Web Search and Data Mining*. ACM New York, USA.

Kohn, A. (1999). From degrading to de-grading. Rev. ed. Boston: Houghton Mifflin.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Milton, O. (2009). *Making Sense of College Grades: Why the Grading System Does Not Work and What Can be Done About It*. San Francisco: Jossey-Bass.

Milton, O., H. R., P., and J. A., E. (1986). *Making Sense of College Grades*. San Francisco: Jossey-Bass.

Petz, G., Karpowicz, M., Frschu, H., Auinger, A., Winkler, S., Schaller, S., and Holzinger, A. (2012). On text preprocessing for opinion mining outside of laboratory environments. In *Active Media Technology*, pages 618–629. Springer.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for idf. In *Journal of Documentation*, volume 60.

Rockoff, J. and Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. In *Labour Economics, Volume 18, Issue 5*, pages 687–696.

Spärck Jones, K. (1972). *A statistical interpretation of term specificity and its application in retrieval*.

Suskie, L. (2009). *Assessing Student Learning: A common sense guide*. Malden: Jossey-Bass A Wiley Imprint, San Francisco, 2nd edition.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley.

Walvoord, B. E. and Anderson, V. J. (2009). *Effective Grading: A Tool for Learning and Assessment in College*. Paperback, 2nd edition.

Wu, H. C., Luk, R. W. P., Wong, K. F., and Kwok, K. L. (2008). Interpreting tf-idf term weights as making relevance decisions. *ACM Trans. Inf. Syst.*, 26(3):13:1–13:37.