Adjusting the weights of assessment elements in the evaluation of Final Year Projects

Mikel Villamañe, Mikel Larrañaga, Ainhoa Alvarez, Begoña Ferrero
Department of Languages and Computer Systems
University of the Basque Country (UPV/EHU), Spain
{mikel.v, mikel.larranaga, ainhoa.alvarez, bego.ferrero}@ehu.eus

ABSTRACT

The authors of this paper have defined a continuous evaluation methodology for Final Year Projects, in which six different evaluable items are involved. However, establishing the weights of each assessment element in the evaluation of Final Year Projects is a complex process, especially when several teachers are involved [3] like in this case. In this paper, the experiment carried out in order to establish the weight each assessment element should have in the final mark of a Final Year Project is described.

Keywords

Final Year Projects, weight adjustment, experts' validation

1. INTRODUCTION

Finishing a Final Year Project (FYP) is a challenging task for all the involved actors, either students or lecturers. In a previous work, the authors conducted a study and concluded that the main problems during projects' development are related to the evaluation process [7].

In many universities, the evaluation of the FYPs has been mainly based on a final dissertation of the work and a public oral defense in front of an examination board. This approach presents several drawbacks [6]. In order to overcome them, a set of 8 experts (teachers from the University of the Basque Country, with more than 10 years supervising FYPs) defined six elements to be taken into account and the responsible for their evaluation.

The supervisor of the project evaluates: an initial report including the project planning and requirements (*Init_Report*), the result of the design phase of the project (*Design*) and the students' attitude during the process (*Attitude).*

The evaluation board evaluates: the final report of the project (*End_Report*), the oral defense (*Defense*) and the complexity of the project (*Complexity*).

To avoid the subjectivity, an evaluation rubric was created for each of the evaluable elements [4].

2. ADJUSTING THE WEIGHTS OF THE EVALUABLES

According to the proposed FYP grading proposal [7], the final grade is computed as the weighted mean of the scores achieved in the assessable elements. Next, the experiment carried out to adjust those weights is described.

2.1 Data Set & Techniques

In order to develop a model to accurately predict the mark of a FYP, a set of graded FYPs, including the final grade provided by the evaluation board using the traditional grading way and the grades for each of the items for those projects, are required.

In this experiment, 32 FYPs were evaluated. The collected data was randomly split into two data sets, *training set*, which contained 2/3 of the collected data, and the *validation set*, entailing the remaining data.

Adjusting the weight to compute the grade as accurate as possible in relation to the grades given by the evaluation board is a regression problem. Therefore, the first technique tested was the linear regression. In this experiment the target variable is the final mark and the features are the 6 items that according to experts should influence the final mark. The objective is to determine to which extent affects each element the final mark.

During this experiment, negative coefficients were inferred (see Table 1, *LRModel)*. In the case of FYP, a negative value is not applicable as the assessable elements refer to aspects the FYP must satisfy, whilst a negative weight would mean that an undesirable or wrong feature is being evaluated. To overcome this problem, non-negativity constraints in the model should be enforced. Therefore, the Lawson-Hanson Non-negative least-squares technique [2] was used in the second phase of the experiment.

Table 1. Weights of each item in the final mark

	Weights						Analysis results	
	Init_Report	**Design**	**Attitude**	**End_Report**	**Defense**	**Complexity**	**Correlations**	**RMSE**
LRModel	0.24	0.18	-0.08	0.37	0.11	0.15	0.95	0.49
NNLSModel1	0.1	0.26	0	0.31	0.19	0.14	0.97	0.31
NNLSModel2	0.25	0.08	0.08	0.46	0.13	0	0.85	0.55
NNLSModel3	0	0	0	0.52	0.32	0.16	0.96	0.35

2.2 Validation Procedure

The validation process consisted in analyzing the extent to which the obtained model fits the data. With this objective, evaluation boards' judgments and the marks obtained using the weights of the different models were compared computing two different metrics: Pearson correlation coefficients [5] and Root-Mean-Squared Error (RMSE) [1].

The admissible error for the model has to be defined taking into account the peculiarities of the process. In this case, according to the experts, it is a common practice to round the grades to 0.5 points intervals, being very unusual to find grades not matching this criterion. For example, grades such as 7 or 7.5 were observed in the training set, whereas intermediate grades similar to 7.2 were not found. Taking this into account, for this experiment 0.5 has been asset as the maximum admissible error.

2.3 Exploratory Analysis and Working Hypothesis

The identified 6 features are considered independent factors for the final score, as they are evaluated in different stages of the FYP process. To determine the new models to compute the final grades of the FYPs, the authors stated the following hypotheses:

- **H1**: The factors identified by the expert board are appropriate predictors for the final grade of the FYPs.
- **H2**: the complexity of the FYPs is implicitly considered in the other evaluable elements.
- **H3**: The evaluation board can infer all the information needed from the *End_report* and the *Defense*.

Considering these starting hypotheses, the following models were defined for this experiment:

- **LRModel**: Model derived using linear regression and considering all the features. (Hypothesis H1)
- **NNLSModel1**: Model derived using the Lawson-Hanson Non-negative least-squares technique and considering all the features. (Hypothesis H1)
- **NNLSModel2:** Model derived using the Lawson-Hanson Non-negative least-squares technique and considering all the features except *Complexity*. (Hypotheses H1 and H2)
- **NNLSModel3**: a model derived using the Lawson-Hanson Non-negative least-squares technique only considering the *End-report*, the *Defense* and the *Complexity*. (Hypothesis H3)

3. RESULTS

In this experiment, the models described above were derived using the *training set* and tested on the *validation* set.

As it can be observed in Table 1, the linear regression technique, used for *LRModel*, led to a model with negative coefficients for some features (*Attitude*). Although the performance was remarkably good, this is not an admissible model to grade FYPs because it would mean that negative aspects of the project are being measured.

NNLSModel2 had an RMSE of 0.55 points, which did not fit in the defined admissible error range. *NNLSModel1* computed grades with 0.97 correlation with the evaluation boards' and 0.31 RMSE, whereas *NNLSModel3* achieved 0.35 RMSE.

Taking into account the calculated RMSE, the best model is *NNLSModel1* where all the features identified by the expert board are used (including *Complexity*). However, in this model *Attitude* has a weight of 0, i.e., it is not a statistically significant predictor for the final mark. Moreover, as shown in Table 1, with *NNLSModel1* an error of 0.31 in a 10-point scale has been achieved. As previously mentioned, this is an admissible error for the evaluation of FYPs because it is inferior to 0.5.

4. CONCLUSIONS AND FUTURE WORK

This paper has presented the experiment carried out in order to adjust the weights of assessment elements for the evaluation of FYPs. Several models have been evaluated, achieving a model with an error of 0.31 in a 10-point scale. One of the main results of the experiment is that the student's attitude *(Attitude)* is not statistically significant to predict the final mark.

The best performing model considers elements that must be evaluated by the supervisor of the FYP in addition to the elements assessed by the evaluation board. This suggests that, even if the evaluation board can give a grade, for a detailed evaluation, the opinion of the person who better knows the project is required.

The main future work is related to the adjustment of weights for each dimension of the rubrics. Additionally, the authors will continue validating the obtained model with new evaluations.

5. ACKNOWLEDGMENTS

This work has been supported by the Basque Government (IT722-13), the Gipuzkoa Council (FA-208/2014-B) and the University of the Basque Country (UFI11/45).

6. REFERENCES

[1] Chai, T. and Draxler, R.R. 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*. 7, 3 (2014), 1247–1250.

[2] Lawson, C. and Hanson, R. 1995. *Solving Least Squares Problems*. Society for Industrial and Applied Mathematics.

[3] Quevedo, J.R. and Montanes, E. 2009. Obtaining Rubric Weights for Assessments by More than One Lecturer Using a Pairwise Learning Model. (Cordoba, Spain, Jul. 2009), 289–298.

[4] Stevens, D.D., Levi, A.J. and Walvoord, B.E. 2012. *Introduction to Rubrics: An Assessment Tool to Save Grading Time, Convey Effective Feedback, and Promote Student Learning*. Stylus Publishing.

[5] Taylor, R. 1990. Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*. 6, 1 (1990), 35–39.

[6] Valderrama, E., Rullan, M., Sánchez, F., Pons, J., Mans, C., Giné, F., Jiménez, L. and Peig, E. 2009. Guidelines for the final year project assessment in engineering. *39th IEEE Frontiers in Education Conference, 2009. FIE '09* (San Antonio, Texas, EE.UU., Oct. 2009), 1–5.

[7] Villamañe, M., Ferrero, B., Álvarez, A., Larrañaga, M., Arruarte, A. and Elorriaga, J.A. 2014. Dealing with common problems in engineering degrees' Final Year Projects. (Madrid, 2014), 2663–2670