E-moderation of Answer-scripts Evaluation for Controlling Intra/Inter Examiner Heterogeneity

Kissan Gauns Dessai
Department of Computer Science
Govt. College of Arts, Science & Commerce
Quepem-Goa, India
kissangd@com

Venkatesh V. Kamat
Department of Computer Science & Technology
Goa University
Taleigao-Goa, India
vvkamat@unigoa.ac.in

Abstract — **Public examinations are conducted worldwide for certification, placement, promotion, etc. As these examinations are high stake examinations, evaluation of the answer-scripts needs to be carried out in a uniform, error-free and unbiased manner. However, the large quantum of answer-scripts pertaining to each subject/course paper invariably introduces evaluation anomalies. Coupled with this, evaluation also suffers from intra/inter examiner heterogeneity and subjectivity. Some of the currently used approaches such as moderation of answer-scripts, in-house verification, personal verification, re-evaluation of answer-scripts and scaling of marks, only provide cursory relief from anomalous and heterogeneous evaluation. This is apparent from alarmingly increasing cases of verification/re-evaluation converging into significant changes in the original marks. In this paper, we propose an E-moderation scheme using machine learning techniques to classify each answer evaluation as negligent or normal and further predict scale.**

Keywords- ***Public Examination; Evaluation Anomalies; Examiner Heterogeneity; E-moderation; Marks Tuning.***

I. INTRODUCTION

Public examinations under the ambit of summative examination are conducted for promotion, placement, certification, and accountability [1]. As public examinations are taken up by large number of examinees, it creates a large quantum of answer-scripts for evaluation. In a public examination system with a large number of answer-scripts pertaining to each course paper/subject, it is not possible to get all the answer-scripts evaluated by one examiner. When more than one examiner evaluate the answer-scripts pertaining to a subject, the subjectivity of the respective examiner creeps into the evaluation. Therefore, there is a need to evolve a procedure to ensure uniformity within the examiners so that the effect of 'examiner subjectivity' or 'examiner variability' is minimized.

Also, as evaluation progresses, the same examiner can increase or decrease the standard of evaluation due to various factors such as improved understanding of the subject, the order of evaluation, time of the day, time constraints, fatigue, etc. The evaluation also suffers from 'Hawk-Dove effect' [2], where some examiners are liberal in evaluation and tend to award more marks. Some examiners are strict and tend to give less marks. This apart there are also instances of gross negligence/ lapses in evaluation as apparent from alarmingly increasing cases of personal verification/re-evaluation converging into significant changes in the original marks.

Some of the methods adopted to reduce examiner subjectivity or variability and the attached evaluation anomalies are moderation of answer-scripts, In-house verification, personal verification and re-evaluation [3].

Contributions: This paper proposes a novel approach for classifying the evaluation of answer-scripts as negligent or normal and further predict the tuned marks in an attempt to control the intra/inter examiner heterogeneity in the evaluation. We propose to build an e-moderation model for detection of negligence in evaluation and predicting the tuned marks with the aid of examination data and the machine learning (ML) techniques.

Outline: The remainder of this paper is structured as follows: Section 2 provides the brief description of evaluation anomalies in public examinations and current measures addressing those anomalies along with the related work. Section 3 describes the mechanism adopted for carrying e-moderation of evaluated answer-scripts. Section 4 provides the research methodology used in assessing the usefulness of the proposed approach. Section 5 evaluates the performance of evaluation classifier and tuned marks predictor. Section 6 draws the conclusions and outlines the future work.

II. BACKGROUND AND RELATED WORK

The current study is based on the public examination environment prevailing in most of the academic institutions in India.

A. Evaluation Anomalies and Countermeasures

The two key issues plaguing the evaluation of subjective answer-scripts pertaining to public examinations are negligent evaluation and intra/inter examiner variation in evaluation. If a particular examiner assigns contrasting marks to the similar answer content of two different examinees, we refer it as 'intra examiner' variation. During the process of evaluation the standard of evaluation seldom remains constant due to the factors such as a large number of answer-scripts for evaluation, perceived pattern in the answer-script content, order of evaluation, time of the day, fatigue, time constraints,

978-1-7281-1143-8/18/$31.00 ©2018 IEEE
DOI 10.1109/T4E.2018.00035

etc. These are some of the factors that lead to 'intra examiner' variation in allotment of marks.

When more than one examiner evaluate the answer-scripts pertaining to a subject, the subjectivity of the respective examiner creeps into the evaluation. Inevitably, therefore, there is a difference in average marks and the range of marks awarded by each examiner. This phenomenon is referred to as 'inter examiner' variation in allotment of marks. The presence of a large number of answer-scripts for evaluation, normally gives rise to inconsistency/errors/negligence irrespective of whether it is a single examiner or multi -examiner evaluation.

Many academic institutions address intra/inter examiner variation in evaluation with the help of moderation of assessed answer-scripts. In this process one subject expert acts as a moderator. Moderator picks up some random sample of answer scripts evaluated by each examiner and evaluate them independently. Examiner 'X' needs to evaluate all the answer-scripts again, if major variations are observed. However, it is observed that having rigorous moderation procedures adds little to accuracy and reliability in evaluation, on the contrary delays the assessment and final grading [4].

Scaling techniques are used in many standard and competitive examinations for controlling inter examiner variation in evaluation. However, scaling to a limited extent, is successful in eliminating the general variation which exists from examiner to examiner, but it is not a solution to solve examiner variability arising from the 'Hawk-Dove effect' (strict/liberal valuation) [5].

The computer-assisted grading using rubrics have been shown to help in solving the examiner variation as it clearly identifies specific criteria to be assessed to achieve objectivity in the assessment [6]. There are various tools for computer assisted evaluation using rubrics such as for semi-automatic grading of programming courses [7], grading of descriptive type examinations [8], essay grading [9], standard summative examination answer-scripts grading [3] along with software's such as moodle (https://moodle.org/) and Blackboard (http://www.blackboard.com) for evaluation of essays, assignments and descriptive questions.

Although such systems take care of error prone tasks, they are not fully effective in controlling anomalies in evaluation and intra/inter examiner variation in evaluation.

Machine learning has proven its worth in a large number of applications in solving classification and regression problems. We in this paper, explore some of the machine learning techniques to address the issue of classification of evaluation anomalies in summative examination.

B. *Related Work*

Machine learning techniques are used extensively to solve problems related to classification (e.g. Predicting whether a tumor is benign or malignant [10]; Filtering emails as "spam" or "ham" [11]; Image Classification [12]; Protein fold recognition [13], Regression (e.g. Predict the future stock price based on current price and crucial market parameters [14], clustering (e.g. Botnet detection [15]).

We now discuss some of the significant work on the use of machine learning techniques in evaluation and grading. In a study conducted by [16] used Naive Bayes algorithm to identify poor performers based on the examinee demography and past performance to enable tutors to take remedial measures at an initial stage of learning. In another study [17] used examinees' previous semester marks, class test grade, seminar performance, assignment performance, general proficiency, attendance in class and lab work to predict the end semester marks. A similar work carried by [18] used rule-based systems to predict examinee performance in an e-learning environment using fuzzy association rules. There are some other studies where several classification algorithms have been applied for classifying examinees into groups such as passing or failing [19]. Ref. [20] used multilayer perceptron topology for predicting the likely performance of a examinee being considered for admission into the university.

In summary, there have been a handful of studies identifying the relationship between the examinee grades and past performance or demography using machine learning techniques. However, no methodologically robust study has directly investigated the effectiveness of machine learning techniques in solving the problem of evaluation anomalies.

III. E-MODERATION OF EVALUATION

In order to assess the quality of evaluation and establish uniformity in evaluation, we propose a solution in the form of e-moderation scheme comprising of two parts:

1) Classification of the evaluations carried by each examiner as negligent or normal.
2) Predicting the tuned marks for each examiner for controlling the intra/inter-examiner variation in evaluation.

The entire process of evaluation of answer-scripts along with the proposed e-moderation is illustrated in Fig. 1.

A. *Detection of Negligent Evaluation*

Examiners are provided with computer assisted evaluation system, for carrying evaluation of the subjective answer-scripts and recording the marks. The evaluation system is also designed to record other essential parameters for classification of the evaluations.

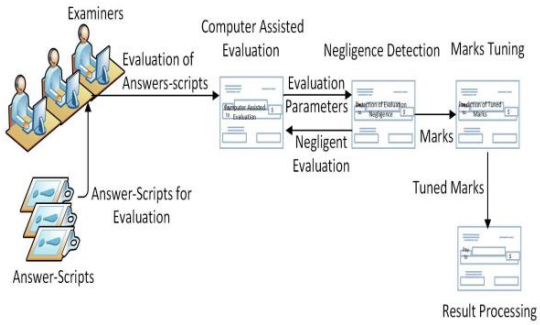

Fig. 1 Evaluation cycle with e-moderation scheme

We take into consideration, the actual time taken for evaluation and time required for evaluation based on examiner reading speed and comprehension accuracy. Similarly, we

consider the actual marks assigned by the examiner for each answer and the maximum marks deserved for it. The given evaluation is considered as negligent, if:

1. There is an apparent variation between the actual time taken for evaluation and examiner reading speed and comprehension accuracy.
2. There is an apparent variation in the marks assigned vis-à-vis number of words contained in the answer.

All those evaluations which fail in the above two tests are classified as negligent. We use an SVM classifier to classify the given evaluations as negligent or normal.

B. Prediction of Tuned Marks

In a typical examination system, multiple examiners are required to carry evaluation of answer-scripts pertaining to a particular subject/course paper. When multiple examiners evaluate the answer-scripts related to a particular subject/course paper, each examiner tends to apply his own yardstick to assess the answer-scripts, resulting in inter examiner variation in evaluation.

The intra/inter examiner variation can be controlled by adjusting the marks assigned by respective subject examiners to one scale adopted by any one examiner. We apply ANN regressor on evaluations carried by different examiners pertaining to a particular course paper and predict the marks as if one examiner had evaluated all those answer-scripts. This way the entire evaluation is normalized onto one common scale to control the 'Hawk-dove effect'.

IV. RESULTS AND DISCUSSION

The current study is undertaken to detect the negligent evaluation and also predict the tuned marks to control intra/inter examiner variation in evaluation.

A. Performance of Evaluation Classifier

We verified the performance of the evaluation classifier using four verification metrics: confusion matrices/classification report, accuracy, ROC and AUC.

The original dataset is split into 70% for training and 30% for testing. The 70% portion of partitioned train data set is further split into a train and a validation partition using 3-fold cross validation technique. The confusion matrices obtained in this process for each of the fold is illustrated in Fig. 2 (a-c). Higher values in the diagonal columns corresponding to the True Negative (TN) and True Positive (TP) as compared to the False Positive (FP) and False Negative (FN) columns, confirms the higher accuracy rate of the model.

We calculated the overall accuracy of the model using an equation:

$$\text{Accuracy} = \frac{TN + TP}{TN + FN + FP + TP} \quad \text{.....(1)}$$

We obtained accuracy of 95% with the simplest linear kernel function. The accuracy of the model further improved to 97% with the use of non-linear kernel (RBF kernel) with C=2 and $\gamma = 1.0$ (refer Fig. 2(d)).

The consistent percentage of 97% for each of precision, recall and f1 produced by the evaluation classifier further confirms the reliability of the model (refer Table. 1).

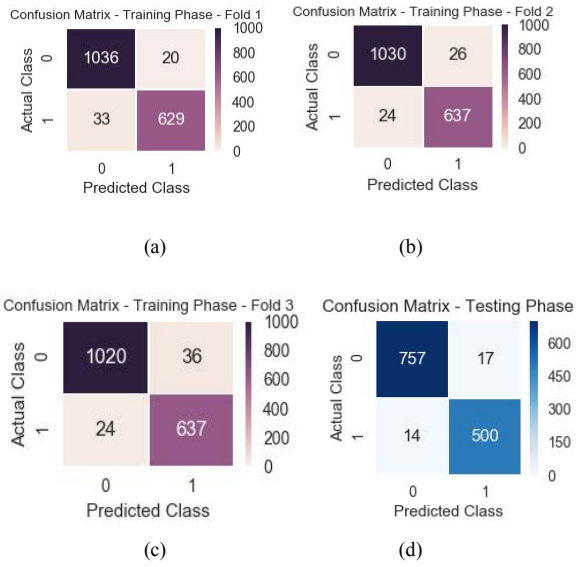

Fig. 2 Heatmap illustrating the confusion matrix for the evaluation classifier. (a) Confusion matrix obtained in the training phase corresponding to fold 1. (b) Confusion matrix obtained in the training phase corresponding to fold 2 (c) Confusion matrix obtained in the training phase corresponding to fold 3. (d) Confusion matrix obtained in the testing phase with the application of the best model of the training phase

Further, we obtained Receiver Operating Characteristic (ROC) curve (Graph of the true positive rate (Sensitivity=$\frac{TP}{TP+FN}$) v/s the false positive rate (Specificity=$\frac{TN}{TN+FP}$) for different cutoff points).

TABLE I. CLASSIFICATION REPORT FOR THE EVALUATION CLASSIFIER

Classification Label	*Precision*	*Recall*	*f1-score*	*Support*
0 – Normal	0.98	0.98	0.98	774
1 – Negligent	0.97	0.97	0.97	514

The high accuracy of the model can be confirmed from the position of the curve following the left hand border and then the top border of the ROC space (refer Fig. 3). The ability of the model to correctly classify negligent and normal evaluation is further confirmed by AUC (Area Under Curve) value of 0.98 (refer Fig. 3).

In addition to the critical factors considered in this study, some other factors such as number of answer-scripts for evaluation, time of the day, time constraints etc. also needs to be explored for further improving the model.

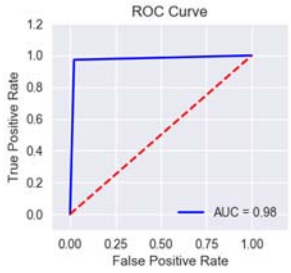

Fig. 3 ROC Curve for the evaluation classifier

B. *Performance of Tuned Marks Predictor*

The evaluation of the answer-scripts was done by four examiners separately, with evaluation carried by each examiner fully blinded from the other examiner to avoid any bias in the evaluation. First, we identified the degree of variation in the actual evaluation of each examiner with the help of ANOVA. An ANOVA test revealed statistically significant differences between the evaluation carried by each of the examiner at the $p < 0.05$ with $F(3,176)=5.568$; sig=0.001. Post-hoc comparisons using the Turkey HSD test also revealed that there is a significant difference in evaluation carried by each examiner at the 0.05 level. We also applied an ANOVA test on marks predicted by our ANN based marks tuner. The ANOVA test corresponding to the marks predicted by our ANN based marks tuner showed insignificant variation at the $p < 0.05$ with $F(3,176)=0.609$; sig=0.610). We also conducted post-hoc comparisons using the Turkey HSD test on marks predicted by ANN based marks tuner. This result also revealed the insignificant variation in the tuned marks predicted with the sig. value ranging from 0.86 to 0.99 at the 0.05 level.

We tried different models for predicting the best outcome, though not a single model produced the best tuned marks for all the examiners involved. We found that different models produced excellent results for different examiners. This means that, there is still scope for improving the proposed model by combining the multiple best models into one optimum model.

V. CONCLUSION AND FUTURE WORK

The evaluation of subjective answer-scripts suffers from the large scale evaluation anomalies and the impact of 'examiner subjectivity' or 'examiner variability'. The currently adopted methods such as moderation and scaling only provide cursory relief from the menace of evaluation anomalies. The current study is undertaken to detect those evaluations that suffer from negligence and also predict the tuned marks in the event of intra/inter examiner variation in evaluation. We used SVM classifier for classifying the evaluation as negligent or normal and used ANN regressor for predicting the tuned marks. Findings from this research indicate that each answer evaluation can be classified into negligent or normal with a great degree of accuracy. This study also provided evidence that we can predict the marks of one examiner based on the evaluation carried by another examiner with the proper training and validation of the model.

REFERENCES

[1] Rovai, A. P. (2000). Online and traditional assessments: what is the difference? The Internet and Higher Education, 3(3), 141-151.

[2] McManus, I. C., Thompson, M., & Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP (UK) clinical examination (PACES) using multi-facet Rasch modelling. BMC Medical Education, 6(1), 42.

[3] Dessai, K. G., Kamat, V. V., & Wagh, R. S. (2014, December). Effective Use of Rubrics in Computer Assisted Subjective Answer-Script Evaluation. Technology for Education (T4E), 2014 IEEE Sixth International Conference on (pp. 95-98).

[4] Bloxham, S. (2009). Marking and moderation in the UK: false assumptions and wasted resources. Assessment & Evaluation in Higher Education, 34(2), 209-220.

[5] Good, F. J., & Cresswell, M. J. (1988). Placing examinees who take differentiated papers on a common grade scale. Educational Research, 30(3), 177-189.

[6] Ahoniemi, Tuukka, and Ville Karavirta (2009). Analyzing the use of a rubric-based grading tool. ACM SIGCSE Bulletin. 41(3). ACM.

[7] Ahoniemi, Tuukka, and Tommi Reinikainen (2006). ALOHA- A grading tool for semi-automatic assessment of mass programming courses. Proceedings of the 6th Baltic Sea conference on Computing education research: Koli Calling 2006. ACM:139-140.

[8] Rane, Archana, Sandip Saha, and M. Sasikumar (2009). A tool for managing descriptive type examinations. International Conference on Management Technology for Educational Practices, July 2009.

[9] Weinberger, A., Dreher, H., Al-Smadi, M., & Guetl, C. (2011). Analytical assessment rubrics to facilitate semi-automated Essay grading and feedback provision. In Proceedings of the Australian Technology Network Assessment Conference 2011 (pp. 170-177). Curtin University.

[10] Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications, 36(2), 3240-3247.

[11] Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. Expert Systems with Applications, 36(7), 10206-10222.

[12] Foody, G. M., & Mathur, A. (2004). A relative evaluation of multiclass image classification by support vector machines. IEEE Transactions on geoscience and remote sensing, 42(6), 1335-1343.

[13] Ding, C. H., & Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics, 17(4), 349-358.

[14] De Fortuny, E. J., De Smedt, T., Martens, D., & Daelemans, W. (2014). Evaluating and understanding text-based stock price prediction models. Information Processing & Management, 50(2), 426-441.

[15] Amini, P., Azmi, R., & Araghizadeh, M. (2014). Botnet Detection using NetFlow and Clustering. Advances in Computer Science: an International Journal, 3(2), 139-149.

[16] Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting Students' Performance in Distance Learning Using Machine Learning Techniques. Applied Artificial Intelligence, 18(5), 411-426.

[17] Bhardwaj, B. K., & Pal, S. (2012). Data Mining: A prediction for performance improvement using classification. arXiv preprint arXiv:1201.3418.

[18] Castro, F., Vellido, A., Nebot, À., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. Evolution of teaching and learning paradigms in intelligent environment, 183-221.

[19] Hämäläinen, W., & Vinni, M. (2006). Comparison of machine learning methods for intelligent tutoring systems. In Intelligent tutoring systems (pp. 525-534). Springer Berlin/Heidelberg.

[20] Oladokun, V. O., Adebanjo, A. T., & Charles-Owaba, O. E. (2008). Predicting students' academic performance using artificial neural network: A case study of an engineering course. The Pacific Journal of Science and Technology, 9(1), 72-79.