A Framework for student engagement level prediction: A user perspective

Abstract: *Most engagement prediction models focus on maximising prediction accuracy in a classification framework or minimising mean square error in a regression framework. However, the goal of an engagement prediction model from a user perspective is to be able to detect disengagement. In this work, we present a regression framework that* encompasses *the approach and evaluation metrics suitable for detecting disengagement. The disengagement level threshold and confidence level threshold are introduced to allow flexibility in the use of the tool. Moreover, the SMOGN oversampling strategy and an undersampling strategy housed in an ensemble are proposed to tackle the imbalance data distribution typically found in engagement detection. They are evaluated on the DAiSEE dataset for benchmarking.*

Keywords: engagement detection, imbalanced regression, oversampling, undersampling

1. Introduction

The COVID-19 pandemic has allowed many to experience the use of technologies for meeting virtually (Venton & Pompano, 2021) (Mukhtar, Javed, Arooj, & Sethi, 2020). However, one of the shortcomings of these virtual meetings are lower engagement levels (Venton & Pompano, 2021). This is particularly salient in virtual learning environments where teachers often feedback that the inability to know if students are engaged in their virtual classrooms unlike in physical classrooms. Hence, the possibility of an engagement level detection tool for teachers is being explored.

Various works (Nezami, et al., 2019) (Abedi & Khan, 2021) (Liao, Liang, & Pan, 2021) (Kaur, Mustafa, Mehta, & Dhall, 2018) (Gupta, Jaiswal, Adhikari, & Balasubramanian, 2016) (Dhall, Sharma, Goecke, & Gedeon, 2020) (Altuwairqi, Jarraya, Allinjawi, & Hammami, 2021) (Whitehill, Serpell, Lin, Foster, & Movellan, 2014) (D'Mello, Dieterle, & Duckworth, 2017) have explored different engagement detection problem setup and methods. Due to nature of data which captures faces of participants, the student engagement datasets are typically kept private and not shared publicly. Out of the many student engagement datasets, two datasets, namely Dataset for Affective States in E-Environments (DAiSEE) (Gupta, Jaiswal, Adhikari, & Balasubramanian, 2016) and Engagement prediction in the Wild (EW) of Emotion Recognition in the Wild (EmotiW) challenge in 2020 (Dhall, Sharma, Goecke, & Gedeon, 2020) are more widely used as these datasets are available upon request. The DAiSEE dataset is annotated by 10 untrained crowdsourcers for every 10 seconds clip while the EmotiW dataset is annotated by 5 experts for every 5 minutes clip. They are both annotated into 4 groups - for DAiSEE dataset, (1) very low, (2) low, (3) high and (4) very high engagement levels; for EmotiW dataset, (0) completely disengaged, (1) barely engaged, (2) engaged, (3) highly engaged. Both datasets have imbalanced distribution with the lack of disengaged samples.

These datasets have been set up as both classification (Abedi & Khan, 2021) (Liao, Liang, & Pan, 2021) and regression (Wu, Yang, Wang, & Hattori, 2020) (Liao, Liang, & Pan, 2021) problems in literature. However, setting up as a regression problem could be more suitable. The closer relationship between (1) very low and (2) low compared to (1) very low and (4) very high engagement can be reflected in the model. Further, the use of a regression model can allow the user to select an engagement level threshold for disengagement, which could differ for different activities.

The main purpose of this engagement tool is to help educators or students themselves detect disengagement. However, due to the imbalanced distribution and sparsity of data, it is difficult to build a model that can accurately predict disengagement, as seen in (Abedi & Khan, 2021) where balancing strategies are required to predict very low / low engagement samples in a classification framework. In this paper, we propose a framework that can help to identify low engagement and provide the user about its confidence level on the prediction in a regression framework. More specifically, we present new metrics that would be useful to the end user and benchmark various strategies to improve performance on imbalanced datasets.

2. Proposed Methodology

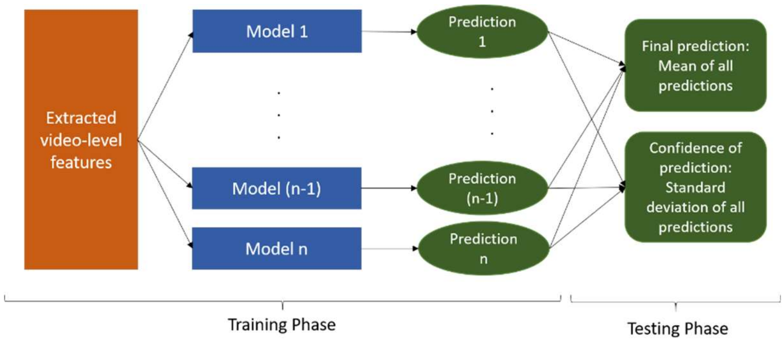

Figure 1. Proposed framework for student engagement level prediction.

The proposed framework for student engagement level prediction is illustrated in Figure 1. From a user perspective, two key factors when using an engagement detection tool are (1) to be able to adjust the disengagement threshold and (2) to know how confident the model is of the prediction made. We refer to this as the flexibility of the engagement detection tool. With a regression model, the value of (1) can be adjusted accordingly with levels of engagement varying from 0 to 1, from disengaged to highly engaged. A lower threshold can be fixed to reduce false positive alerts on disengagement while a higher threshold can be used to filter out more instances of disengagement for retrospective analysis.

First, video-level features would be extracted from the video frames in the dataset (Section 2.1). With the extracted features, multiple models would be built and trained to make multiple predictions. Thereafter, the mean of all the predictions would be calculated as the final prediction and the standard deviation of the predictions would be used as an indicator for the level of confidence for the prediction. The use of a confidence metric helps to address (2), more details are given in Section 2.2. Two varying strategies to manage the imbalance distribution and evaluation metrics would be discussed in Section 2.3 and 2.4 respectively.

2.1. Video-level features

The video-level feature vector is adopted from (Abedi & Khan, 2021), made up of affect and behavioural features. The affect features include the mean and standard deviation of valence and arousal from the consecutive frames using pre-trained Emonet (Abdul-Mageed & Ungar, 2017) on AffectNet (Mollahosseini, Hasani, & Mahoor, 2019), which is one of the largest datasets for facial expressions in the wild. The behavioural features showing the eye and head movement in the video are obtained from OpenFace (Baltrusaitis, Zadeh, Lim, & Morency, 2018), a publicly available toolkit for facial behaviour analysis. The features extracted from OpenFace are the coordinates of gaze position, head position, roll, pitch and yaw for head rotation. With the extracted raw features, we calculated the velocity and acceleration to better represent the gaze and head movement. The trend of gaze and head movement is captured by the mean and standard deviations of velocity and acceleration. We further extended the feature vectors by 5 dimensions using different thresholds on AU45_r (a feature from OpenFace) to determine the number of blinks per frame (Ranti, Jones, Klin, & Shultz, 2020). Following the method described, we constructed the feature vectors for DAiSEE dataset. The overall structure of the 42-element feature vector can be shown in Table 1.

Table 1. Feature Vector Structure

Feature type	Feature description	Dim	Toolkit
Affect	valence mean	1	Emonet
	valence stdv	1	

	arousal mean	1	
	arousal stdv	1	
Behavioural	gaze velocity mean (x,y)	2	OpenFace
	gaze velocity stdv (x,y)	2	
	gaze acceleration mean (x,y)	2	
	gaze acceleration stdv (x,y)	2	
	head translational velocity mean (x,y,z)	3	
	head translational velocity stdv (x,y,z)	3	
	head translational acceleration mean (x,y,z)	3	
	head translational acceleration stdv (x,y,z)	3	
	head rotational velocity stdv (x,y,z)	3	
	head rotational acceleration mean (x,y,z)	3	
	head rotational acceleration stdv (x,y,z)	3	
	blink rate (AU45_c)	1	
Extended behavioural	AU45_r	5	OpenFace

2.2. Ensemble methods for confidence

The final prediction of the approach is calculated as:

$$\hat{y}_{mean} = \frac{\sum \hat{y}_i}{n} \qquad i = 1, \dots, n$$

where $\hat{y}_{mean}$ is the final prediction, $\hat{y}_i$ is the prediction of model i and n is the total number of models. By setting a disengagement threshold θ_{thres}, where $\hat{y}_{mean} \leq \theta_{thres}$, one would be able to identify disengaged samples.

The confidence (standard deviation) of the model prediction is calculated as:

$$\hat{y}_{SD} = -\sqrt{\frac{\sum(\hat{y}_i - \hat{y}_{mean})^2}{n}}$$

where $\hat{y}_{SD}$ is the confidence of the model prediction. The higher the value, the more confident is the model. By using $\hat{y}_{SD}$, samples that have diverse predictions from different models (lower confidence levels) can be identified and dropped using a standard deviation threshold, σ_{thre} .

2.3. Strategies for Imbalanced Dataset

In engagement detection, we hope to capture instances where students have low engagement, to either bring students back to attention or to perform a retrospective analysis for understanding. Therefore, the important ranges are towards the lower range of the engagement level scale, and these are poorly represented in the typical engagement detection datasets. To tackle this, two different strategies are proposed.

2.3.1. Data-based strategy - SMOGN: Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise

SMOGN is a popular strategy for tackling imbalanced regression problems (Branco, Torgo, & Ribeiro, 2017). It combines random under-sampling strategy with two over-sampling strategies: SMOTER (Torgo, Ribeiro, Pfahringer, & Branco, 2013) and the introduction of Gaussian Noise. This combination allows the generation of synthetic examples by reducing the risks faced by SMOTER and at the same time greater diversification. This helps reduces the imbalance between the classes and sample each class equitably. More details can be found in (Branco, Torgo, & Ribeiro, 2017).

2.3.2. Model-based strategy - Undersampling for each model in ensemble

In this strategy, each model in an ensemble would be built and trained using only a subset of the whole dataset (undersampling). The number of samples in each class can be decided by the user to allow for a more balanced dataset. Each model would be trained using a different subset, randomly selected from the whole dataset. This would vary the predictions in the various models. Hence, if the different models have similar predictions, the final prediction $\hat{y}_{mean}$ would be considered to have higher confidence.

2.4. Evaluation Metrics

In most regression works, the metric used for evaluation is the Mean Square Error (MSE). While MSE is useful for comparing models against one another, it does not provide insights that the user could relate. We propose a few other metrics to help us further evaluate the model's effectiveness on engagement detection. We separate the MSE into MSE for low engagement samples and high engagement samples, namely Low Engagement Mean Square Error (LE_MSE) and High Engagement Mean Square Error (HE_MSE) respectively. This helps us to evaluate the model's effectiveness in high and low engagement samples and is particularly useful in a very imbalanced dataset. Moreover, we introduce the calculation of true positive (TP), false positive (FP) count and F1 score when given a certain θ_{thres}.

3. Experimental Study

In our experiments, we demonstrated the data-based strategy of SMOGN and the model-based strategy of undersampling on the DAiSEE (Gupta, Jaiswal, Adhikari, & Balasubramanian, 2016) datasets.

For the SMOGN strategy, Random Forest (RF) regressor, Gradient Boosting Regressor (GBR), support vector regression (SVR) and multi-layer perceptron (MLP) models were used. For RF and GBR, the total number of estimators is 100. For SVR, the epsilon value is set at 0.01. For MLP, we train and test on a 3 layered MLP model having 64 hidden neurons and tanh activation function in each layer. This is followed by a final sigmoid layer to calculate the regression scores. We use the Adam optimizer with MAE loss with batch size 32 for 200 epochs to train the model. The setting for the SMOGN is to over-sample the samples with lower distribution. In DAiSEE, we oversampled (1) very low and (2) low samples.

For the undersampling ensemble approach, a hundred decision trees with varying number of samples per class were tested - three settings were used: (1) balanced number of samples per class, (2) 6 times more samples selected in non-minority classes and (3) distribution similar to original distribution of data. θ_{thres} is set to 0.5 in our experiments.

3.1. Datasets

The original DAiSEE dataset that was downloaded is split into 3 phases, containing 5482 clips from 33 users for training, 1720 clips from 38 users for validation, and 1866 clips from 32 users for testing. The training and validation sets were combined for training in our experiments. A total of 143 clips that were without labels were removed from the test set. Table 2 provides the distribution of samples across the train and test sets in the DAiSEE dataset.

Table 2. DAiSEE Data Distribution

Phase	0	1	2	3	Total
Train	57	374	3561	3210	7202
Test	4	81	861	777	1723
Total	61	455	4422	3987	8925

3.2. Experimental Results

Table 3. DAiSEE dataset results

Approach	No.of models	MSE	LE_MSE	HE_MSE	TP (count)	FP (count)	F1 score
RF	1(100)	0.0224	0.1016	0.0183	0	1	0
GBR	1(100)	0.022	0.1078	0.0175	2	2	0.045
SVR	1	0.0271	0.1039	0.0231	3	9	0.062
MLP	1	0.0336	0.1072	0.0297	0	0	0
Data-based strategy: SMOGN							
RF(SMOGN)	1(100)	0.0258	0.0849	0.0228	4	8	0.082
GBR(SMOGN)	1(100)	0.0373	0.0583	0.0362	21	119	0.187
SVR(SMOGN)	1	0.04	0.0793	0.0379	13	142	0.108
MLP(SMOGN)	1	0.052	0.0605	0.0516	33	350	0.141
Model-based strategy: Undersampling each model in ensemble							
Decision Trees (50, 50, 50, 50)	100	0.0594	0.0258	0.0611	55	463	0.182
Decision Trees (50, 300, 300, 300)	100	0.0366	0.0446	0.0362	27	136	0.218
Decision Trees (50, 100, 300, 300)	100	0.0288	0.0615	0.0271	21	70	0.239

The results for the DAiSEE dataset are presented in Table 3. Without any strategies, RF and GBR have achieved better MSE than SVR and MLP. However, when looking at the TP and FP counts, GBR would be a more suitable model as it can detect 2 instances of student being disengaged compared to 0 in RF. When the imbalance data strategies are employed, there is a trade-off of HE_MSE for a lower LE_MSE. This in turn allowed more TP counts as LE samples are more accurate, as well as higher FP counts when HE samples are less accurate. In the model-based strategy, the MSE is lowest when the ratio is the samples is similar to the current distribution of data among classes.

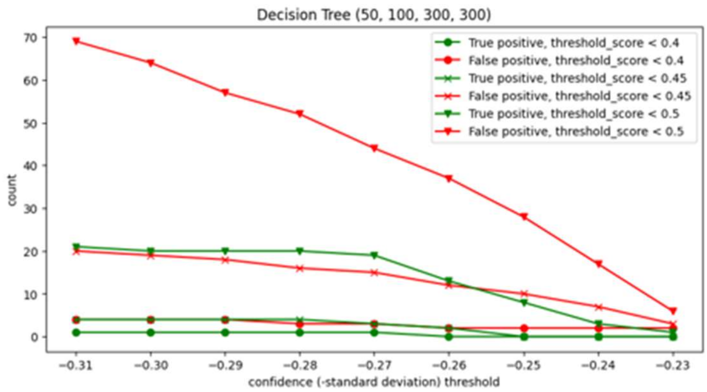

Figure 2. Variation of θ_{thres} and σ_{thres} in ensemble approach

Figure 2 presents the changes in TP and FP, when θ_{thres} and σ_{thres} are varied in Decision Trees (50, 100, 300, 300) approach. Given $\theta_{thres} = 0.5$, the TP and FP can be reduced to 20 and 52 respectively when the $\sigma_{thres} = -0.28$.

4. Conclusion and Future Directions

This work presents a framework that focuses on building and evaluating an engagement detection tool that can detect disengagement. The framework allows flexibility with the introduction of the disengagement level threshold θ_{thres} and confidence level threshold σ_{thres}. Furthermore, the work evaluated the SMOGN oversampling strategy and an undersampling strategy housed in an ensemble on the DAiSEE dataset. The results reflected the challenges of the engagement detection problem - specifically in identifying disengagement. Our future work would explore effective approaches on the collection and annotation of data for the engagement detection problem to allow us to build a useful engagement detection tool for educators.

References

Abdul-Mageed, M., & Ungar, L. (2017). {E}mo{N}et: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 718-728). Vancouver, Canada: Association for Computational Linguistics.

Abedi, A., & Khan, S. S. (2021). Affect-driven Engagement Measurement from Videos. *ArXiv.*

Abedi, A., & Khan, S. S. (2021). Improving state-of-the-art in Detecting Student Engagement with Resnet and TCN Hybrid Network. *18th Conference on Robots and Vision (CRV),* (pp. 151-157).

Altuwairqi, K., Jarraya, S. K., Allinjawi, A. A., & Hammami, M. (2021). Student behavior analysis to measure engagement levels in online learning environments. *Signal, Image and Video Processing*, 1387-1395.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. 2018 *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, (pp. 59-66).

Branco, P., Ribeiro, R. P., & Torgo, L. (2016). {UBL:} an {R} package for Utility-based Learning. *arXiv.*

Branco, P., Torgo, L., & Ribeiro, R. P. (2017). SMOGN: a pre-processing approach for imbalanced regression. *First international workshop on learning with imbalanced domains: Theory and applications*, (pp. 36-50).

Dhall, A., Sharma, G., Goecke, R., & Gedeon, T. (2020). EmotiW 2020: Driver Gaze, Group Emotion, Student Engagement and Physiological Signal Based Challenges. *Proceedings of the 2020 International Conference on Multimodal Interaction* (pp. 784-789). Virtual Event, Netherlands: Association for Computing Machinery.

D'Mello, S., Dieterle, E., & Duckworth, A. (2017). Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational psychologist*, 104-123.

Gupta, A., Jaiswal, R., Adhikari, S., & Balasubramanian, V. (2016). DAISEE: Dataset for Affective States in E-Learning Environments. *arXiv*.

Kaur, A., Mustafa, A., Mehta, L., & Dhall, A. (2018). Prediction and Localization of Student Engagement in the Wild. *Digital Image Computing: Techniques and Applications (DICTA)*, 1-8.

Liao, J., Liang, Y., & Pan, J. (2021). Deep Facial Spatiotemporal Network for Engagement Prediction in Online Learning. *Applied Intelligence*, 6609–6621.

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, 18-31.

Mukhtar, K., Javed, K., Arooj, M., & Sethi, A. (2020). Advantages, Limitations and Recommendations for online learning during COVID-19 pandemic era. *Pakistan journal of medical sciences*, S27.

Nezami, O. M., Dras, M., Hamey, L., Richards, D., Wan, S., & Paris, C. (2019). Automatic Recognition of Student Engagement Using Deep Learning and Facial Expression. *ECML/PKDD*.

Ranti, C., Jones, W., Klin, A., & Shultz, S. (2020). Blink Rate Patterns Provide a Reliable Measure of Individual Engagement with Scene Content. *Scientific Reports.*

Torgo, L., Ribeiro, R. P., Pfahringer, B., & Branco, P. (2013). SMOTE for Regression. *EPIA 2013:Progress in Artificial Intelligence*, (pp. 378-389).

Venton, B., & Pompano, R. R. (2021). Strategies for enhancing remote student engagement through active learning. *Analytical and Bioanalytical Chemistry*, 1507-1512.

Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., & Movellan, J. R. (2014). The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions. *IEEE Transactions on Affective Computing*, 86-98.

Wu, J., Yang, B., Wang, Y., & Hattori, G. (2020). Advanced Multi-Instance Learning Method with Multi-Features Engineering and Conservative Optimization for Engagement Intensity Prediction. *Proceedings of the 2020 International Conference on Multimodal Interaction* (pp. 777-783). Virtual Event, Netherlands: Association for Computing Machinery.