Compare Several Supervised Machine Learning Methods in Detecting Aberrant Response Pattern

Yi Lu, Yu Zhang, Lorin Mueller

Federation of State Boards of Physical Therapy

Abstract

An aberrant response pattern, e.g., a test taker is able to answer difficult questions correctly, but is unable to answer easy questions correctly, are first identified lz and lz*. We then compared the performance of five supervised machine learning methods in detecting aberrant response pattern identified by lz or lz*.

1 Introduction

Investigating fraudulent testing behavior, especially for high-stakes assessments, has been a common practice for maintaining test score validity. In practical assessment, one of the important problems is to ensure that the test taker's response pattern is consistent with the expected item score pattern. When the difference between the observed and the expected pattern is large, it is classified as an aberrant response pattern (Magis, Raiche, & Beland, 2012; Meijer & Tendeiro, 2014). One example is test taker is able to answer difficult questions correctly, but is unable to answer easy questions correctly. Lz and its modification Lz*, two well-known person-fit statistics are applied in the study to detect aberrant response pattern specified above.

The rapid advancement of machine learning (ML) techniques has led to their widespread application across various domains. In recent years, several studies have conducted comprehensive comparisons of machine learning models to understand their relative strengths and limitations across diverse tasks (e.g., Caruana and Niculescu-Mizil, 2006; Neagu et al., 2007; Raschka, 2018). Collectively, these studies provide a foundational basis for applying and evaluating machine learning algorithms in the present study, which focuses on detecting aberrant response patterns using indices such as the lz and lz* statistics. In the field of educational science, several studies explored machine learning to detect exam cheating (e.g., Man et al., 2019; Pan et al., 2022; Zopluoglu, 2019). There are relatively few studies implementing machine learning methods to investigate aberrant response pattern as specified in the current study.

2 Data

Data used for this study was selected from a licensure exam that is administered multiple times each year. We selected one test form that was administered twice in one year for this study. We used item responses from 2561 examinees who took this form in April as training data. We used item responses from 492 examinees who took the same form in October as test data. There were 200 scored items in this form. Item response for these 200 items was taken as input features. The target variable for each examinee is either flagged as an aberrant response pattern or not based on lz or lz* person fit statistics. In literature, the cutoff value of -4 is used to flag examinees of aberrant response patterns (Tendeiro, Meijer, & Niessen, 2016). In our operational analysis, we used the criteria listed in Table 1 on page 7 to flag aberrant response pattern. Using flagging criteria in Table 1, "flagged #" column in Table 2 on page 7 lists the number of flagged cases in training and test data based on lz and lz* indices, respectively. For our data, the examinees with aberrant response pattern are the minority. A much smaller number of positive cases (aberrant response pattern examinees) can lead to bias in model prediction. To handle the issue of data imbalance, we then conducted data simulation. That is, based on the response pattern of the flagged cases, we simulated one time and two times of examinees that have very similar responses as the flagged aberrant response pattern. The last two columns in Table 2 present the simulated number of aberrant response pattern. Those simulated cases were then randomly inserted and replaced normal response

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 21–24

pattern in the original data. In this way, the total number of examinees in training and test data remain the same.

3 Methods

3.1 Lz and Lz* Person-fit Statistics

Drasgow, Levine, & Williams (1985) proposed a standardized version of lz

$$lz = \frac{l_0 - E(l_0)}{V(l_0)} \quad (1)$$

Where l_0 is the log likelihood function of any response pattern, $E(l_0)$ and $V(l_0)$ are the mean and variance of l_0

Snijder (2001), proposed lz*, in which true ability estimates were replaced by sample ability estimates. Magis et al (2012) illustrated lz* as

$$lz^* = \frac{Wn(\hat{\theta}) - c_n(\hat{\theta}) * r_0(\hat{\theta})}{\tilde{V}[l_0(\hat{\theta})]^{1/2}} \quad (2)$$

where $Wn(\theta)$ is a statistic, $r_0(\theta)$ is an estimator, $c_n(\theta)$ is a function modifying $r_0(\theta)$, $\tilde{V}[l_0(\theta)]$ is the modified variance. Magis et al. (2012) has detailed illustrations of those statistics. From equations 1 and 2, we can say that lz* index is a rescaled version of lz by adjusting both its mean and its variance. Lz and lz* are implemented in the current study to identify aberrant response patterns, as illustrated in the data section.

.

3.2 Supervised Machine Learning Methods

Machine learning is broadly categorized into four main types: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. As stated below, five supervised learning methods are implemented in the current study to flag aberrant response pattern identified by lz or lz*.

K-Nearest Neighbor (KNN): KNN is a learning algorithm that attempts to classify new samples by allocating them to the class of the most similar labeled cases. In this study, the KNN algorithm was employed to classify examinee response vectors flagged by the lz or lz* indices as either aberrant or normal. The algorithm does not make assumptions about the underlying data distribution, making it particularly suitable for exploratory and diagnostic contexts. The simplicity and interpretability of KNN provide a valuable benchmark against which more complex models—such as neural networks or Support Vector Machines—can be compared.

Naïve Bayes: The Naïve Bayes classifier is a probabilistic machine learning model based on Bayes' Theorem. Bayes' Theorem is formally expressed as:

$$P(C_k|x) = \frac{P(x|C_k) * P(C_k)}{P(x)} \quad (3)$$

Under the naïve conditional independence assumption, the joint likelihood simplifies to a product of individual feature likelihoods:

$$P(C_k|x_1, x_2, \ldots x_n) \propto P(C_k) \prod_{i=1}^{n} \quad P(x_i|C_k) \quad (4)$$

The classification rule then becomes:

$$\hat{y} = \underset{k \in \{1,k\}}{arg\ max}\ P(C_k) \prod_{i=1}^{n} \quad P(x_i|C_k) \quad (5)$$

Based on equation above, Naïve Baye classification algorithm can be used for categorizing new observation into predefined classes for the initiated data. In this study, the Gaussian Naïve Bayes variant was applied to detect aberrant response pattern identified by the lz or lz* indices. The model was implemented using the GaussianNB class from the sklearn.naive_bayes module in Python.

Logistic regression: Logistic regression models the probability that a given input belongs to a specific class. It does this by applying the sigmoid (logistic) function to a linear combination of the input features (Hosmer, Lemeshow, & Sturdivant, 2013).

The sigmoid function is defined as:

$$S(y) = \frac{1}{1+e^{-y}} \quad (6)$$

In the context of logistic regression, the input to the sigmoid function is a linear combination of the predictor variables:

$$p = \frac{1}{1+e^{-(mx+b)}} \quad (7)$$

Where *p* is the estimated probability that the instance belongs to class 1 (e.g., exhibiting aberrant response pattern), *m* represents the weight coefficients (slopes), *X* is the feature vector (e.g., item responses), and *b* is the intercept (bias).

Logistic regression learns these parameters during model training by maximizing the likelihood of the observed data. In binary classification, a threshold (typically 0.5) is applied to the predicted probability to assign class labels. The model was implementedusing the LogisticRegression class from the sklearn.linear model module in Python.

Support Vector Machine (SVM): The central idea behind SVM is to find the optimal hyperplane that best separates data points from different classes in a high-dimensional space. For binary classification, as in the current study, the goal is to maximize the margin between the two classes—the distance between the hyperplane and the nearest data points from each class, known as support vectors.

In this study, a Support Vector Machine (SVM) classifier was employed to detect examinees with aberrant response patterns, as flagged by the lz or lz* indices. The SVM model was implemented using the SVC class from the scikit-learn library in Python. The default SVM configuration with a radial basis function (RBF) kernel was used, which allows the model to capture non-linear relationships in the data.

Neural networks (NNs): NNs are a class of machine learning models inspired by the structure and function of the human brain. They consist of layers of interconnected processing nodes (neurons), where each neuron applies a transformation to the input and passes the result to subsequent layers. Each connection between neurons is associated with a weight that is learned during training through optimization algorithms such as stochastic gradient descent and backpropagation. To classify examinees based on aberrant response pattern identified by the lz or lz* indices, a feedforward neural network was implemented using TensorFlow and Keras.

In the current study, the architecture of the neural network included the following items:

- An input layer with 200 features (corresponding to the number of items),
- Two hidden layers with ReLU activation functions,
- Dropout layers for regularization to mitigate overfitting, and
- A final output layer with a sigmoid activation function for binary classification.

4. Software for Estimation

In this experimental stage, we used Google Colab for estimation. Oversample method was applied in Colab to make sure all aberrant response patterns have been sampled when training the model.

5. Results

One essential tool to evaluate the performance of machine learning models is confusion matrix. A confusion matrix is a simple table that shows how well a classification model is performed by comparing its predictions to the actual results. A confusion matrix adapted to the context of the current study is presented in Table 3 on page 7. Below is a brief explanation on evaluation metrics that applied in the study to evaluate the performance of these supervised machine learning mothods.

$$\text{Precision}=\frac{TP}{TP+FP}$$

Precision focuses on the accuracy of the model's positive predictions. It tells us how many of the instances predicted as positive are actually positive.

$$\text{Recall/Sensitivity} = \frac{TP}{TP+FN}$$

Recall measures the proportion of correctly predicted positive instances among all actual positive instances.

$$\text{F1score}=2 * \frac{Precison * Sensitivity}{Precision + Sensitivity}$$

F1 score combines precision and recall into a single metric to balance their trade-off. It provides a better sense of a model's overall performance, particularly for imbalanced datasets. F1 score ranges from 0 to 1, with 1 indicating the best possible performance.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FN+FP}$$

Accuracy measures how often the model's predictions are correct overall. It gives a general idea of how well the model is performing.

In the current study, under different conditions on the number of aberrant response pattern, the resulting classification performance was compared among five supervised machine learning models. Tables 4 and 5 on pages 8 and 9 summarize the classification performance of five machine learning models in detecting aberrant response patterns as identified by the Lz and lz* index, respectively.

Results in these two tables show that, under the condition of the real number of flagged cases, most models—particularly KNN and SVM—struggled to detect aberrant responses, often yielding near-zero F1-scores. Logistic regression consistently achieved high precision but suffered from low recall, while Naïve Bayes and neural networks offered more balanced but modest performance. These results underscore the effectiveness of simulation-based data

augmentation for enhancing model sensitivity and suggest that sample size and class balance are critical factors in building reliable aberrant response detectors.

6. Conclusion

In this study, we implemented five supervised machine learning models in detecting aberrant response pattern identified by lz and lz* indices. Across both the Lz and Lz* indices, machine learning models demonstrated consistently high accuracy in identifying normal response patterns. However, performance in detecting aberrant response patterns varied considerably and was highly sensitive to class imbalance. As the number of aberrant responses increased through simulation (1x and 2x the original cases), all models showed marked improvement in identifying aberrant patterns, with F1-scores for class 1 increasing by 2–3 times or more.

In our research, the primary goal of this study has been to compare and choose the best machine learning models. Based on the evaluation metrics—including precision, recall, and F1 score—logistic regression and neural network models demonstrated the strongest performance in detecting aberrant response patterns (in the condition of a real number of aberrant response pattern). However, it is important to note that training the neural network required substantially longer computation time compared to logistic regression. While both models show promise, their effectiveness should be further validated using independent datasets to ensure generalizability. Future research may also explore the potential of alternative machine learning models to enhance detection accuracy and efficiency in various operational contexts.

References

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning*, 161–168.

Drasgow, F., Levine, M. V., & Williams, M. E. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67–86.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of Snijders's l_z^* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics, 37* (1), 57–81.

Man, K., Harring, J. R., & Sinharay, S. (2019). Use of Data Mining Methods to Detect Test Fraud. *Journal of Educational Measurement*, 56(2), 251–279.

Meijer, R.R. & Tendeiro, J. N. (2014). The use of person-fit scores in high-stakes educational testing: How to use them and what they tell us. Law School Admission Council Research Report 14-03, March 2014.

Neagu, D. C., Guo, G., Trundle, P. R., & Cronin, M. T. D. (2007). A Comparative Study of Machine Learning Algorithms Applied to Predictive Toxicology Data Mining. *Journal of Chemical Information and Modeling*, 47(2), 716–729.

Pan, Y., Sinharay, S., Livne, O., & Wollack, J. A. (2022). A Machine Learning Approach for Detecting Item Compromise and Preknowledge in Computerized Adaptive Testing. *Psychological Test and Assessment Modeling*, 64(4), 385–424.

Raschka, S. (2018). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. arXiv preprint arXiv:1811.12808.

Snijders, T. A. B. (2001). Asymptotic Null Distribution of Person Fit Statistics with Estimated Person Parameter. *Psychometrika, 66*(3), 331–342.

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). PerFit: An R Package for Person Fit Analysis in IRT. *Journal of Statistical Software, 74*(5). doi: 10.18637/jss.v074.i05

Zopluoglu, Z. (2019). Detecting Examinees with Item Preknowledge in Large-Scale Testing Using Extreme Gradient Boosting (XGBoost). *Educational and Psychological.*