Introduction

AI is a powerful tool with a variety of applications. One of the most popular uses for AI is data analysis. An AI model is capable of analyzing large datasets and making predictions with remarkable accuracy. However, with smaller datasets these models may jump to inaccurate conclusions based on the few data points available. This behavior is called overfit and can be a serious problem when there is a limited supply of data. As a part of the DREU research program, we investigated the potential of different types of models to correctly classify types of student anxiety from a small dataset.

Literature Review

A typical AI model may run into any number of problems on a small dataset, including overfit. There is a lot of previous research into data overfit that provides insight into this problem. Overfit occurs when an AI model is more complex than it needs to be, either because irrelevant parameters were used (for example, using awareness of anxiety resources to predict type of anxiety) or because the model chosen was too complex (for example, using a neural net for a linear regression problem) [1]. Our research focused on the latter cause of overfit. Although all model types can be affected by overfit, this problem affects different models differently; For example, support vector machine (SVM) models were designed to (among other things) reduce overfit compared to other models, but a poorly selected error penalty can cause these models to overfit as well [2].

Of course, overfit is far from the only problem that may plague an AI model trained on a small dataset. As such, we researched some of the models that are considered useful for small datasets. Random forest models are effective at classifying data with small datasets [3], and they rarely suffer from overfit [4]. Additionally, rule-based models, a type of AI that uses domain knowledge rather than machine learning to classify data, are not dependent on a training dataset [5], and as such do not care if the dataset is small or imbalanced. All of these models may be of interest as effective classifiers for small datasets.

Methodology

We used a dataset of survey results from international graduate students at Florida Tech for our research. This survey contained free responses in addition to the multiple-choice questions, so it had to be manually parsed for spelling errors, delimited for machine readability, and filtered so that unique results that only appeared once in the dataset would be recognized as such. Each respondent was also classified as having more “academic” or “personal” anxiety, if one of these seemed definitive enough to use as ground truth, or as “unsure” if this was not the case. Although these “unsure” results were not included in the training set, their results were predicted as academic, personal, or close enough to the threshold (for those classifiers that had a clear threshold) that the result is, in practice, indeterminate. There were 39 total entries in the dataset, and only five of them were initially classified as having more academic anxiety.

We then created AI models of the four types discussed in Literature Review (Linear Classifier, SVM, Random Forest, and Rule-Based) and tested their ability to classify the entries in the dataset. The machine learning models (all except Rule-Based) were implemented using SciKitLearn's machine learning packages, whereas the rule-based model was created from scratch based on our knowledge of the domain and analysis of patterns within the dataset. This rule-based model incorporated some aspects of "fuzzy" rule-based systems, namely, the assignment of numerical variables based on the data and the use of these variables to determine a result and a relative certainty. Fuzzy sets were not formally implemented because this simpler approach seemed more appropriate for our data, and our results indicated that this method was sufficient.

The machine learning models were trained three times with different subsets of the training data set aside for validation, and the results were examined side-by-side to check for consistency within each model. No such experiment was required for the rule-based model, as the entire dataset was available for validation. We intended to formally compare these models based on accuracy and consistency, but as the results came in it became clear that this was unnecessary.

Results

Note: For readability, our evaluation of each data point as academic/personal/unsure is written in quotes, and the models' evaluations are not. For example, "The AI misclassified one of the "academic" data as personal".

Our first attempt at a linear regression model suffered from severe overfit, and although simplifying the model helped to reduce this it was still very volatile and results varied drastically based on the test set selection. However, this model still fared better than the SVM or random forest model. All three SVM models classified everything in the test set, all "unsure" data, and even some of the "academic" data in the training set as personal, indicating a severe and consistent bias in that direction. The random forest did a little better in some regards, as it at least classified some of the "unsure" data as academic, but it misclassified the "academic" test data as personal in all three cases, indicating that it too had some significant bias. As such, we determined that these models were not effective on this particular dataset, likely because the large ratio of "personal" to "academic" data made it more effective to assume most of the results were personal.

The rule-based classifier fared much better than the others, misclassifying only one of the "academic" and "personal" data and meeting our expectations for all but three of the "unsure" data. All four of these unexpected results were predicted as academic but classified as personal, indicating that even the rule-based model may have had a slight bias in that direction, or that our evaluations were biased to predict academic over personal. The differences seemed to mostly arise from the additional value a human evaluator added to certain qualities (such as talking more in-depth about one cause of anxiety while simply listing off others), which were lost behind the sheer number of personal causes that were mentioned. Creating a rule-based model that can account for subtle qualities such as these may be a topic for future research.

Discussion

Based on this experiment, it seems that the rule-based model is more adept at classifying data from a small unbalanced dataset than any of the other models tested, and it's probably more effective than other machine learning models as well. The obvious reason for this is because the rule-based model can be created independent of the dataset, so a shortage of data has no impact on the model.

However, there are some imperfections in our process that need to be addressed as well. First, the rule-based model was created with the knowledge of how the dataset was classified, meaning the factors that influenced the initial classification were all but guaranteed to be considered the same way in the rule-based model. Additionally, although we have become very familiar with the dataset and claim a thorough understanding of it, we cannot claim to be experts in the field of psychology and human anxiety, and as such our subjective classification of the data could have been erroneous. That being said, the former issue can be considered a benefit of rule-based models as long as the domain knowledge is sound and consistent, and the latter is unlikely to have caused the volatility and overfit of the linear model or the intense bias of the other two. As such, we can solidly conclude that rule-based models are an accurate and consistent alternative to machine learning models that can easily handle a sparse dataset.

Acknowledgements

I would like to thank Dr. Sneha Sudhakaran for her mentorship, leadership, and guidance throughout this research, and the Computing Research Association for facilitating this research through the DREU program.

Works Cited

[1] Douglas M. Hawkins, “The Problem of Overfitting”, *Journal of Chemical Information and Computer Sciences 2004, 44(1), 1-12.* doi: 10.1021/ci0342472 https://pubs.acs.org/doi/full/10.1021/ci0342472

[2] Xiangying Wang and Yixin Zhong, "Statistical learning theory and state of the art in SVM," *The Second IEEE International Conference on Cognitive Informatics, 2003. Proceedings.*, London, UK, 2003, pp. 55-59, doi: 10.1109/COGINF.2003.1225953. https://ieeexplore.ieee.org/abstract/document/1225953

[3] Cha G-W, Moon HJ, Kim Y-M, Hong W-H, Hwang J-H, Park W-J, Kim Y-C. Development of a Prediction Model for Demolition Waste Generation Using a Random Forest Algorithm Based on Small DataSets. *International Journal of Environmental Research and Public Health*. 2020; 17(19):6997. https://doi.org/10.3390/ijerph17196997

[4] T. Hastie et al., *The Elements of Statistical Learning, Second Edition*,

DOI: 10.1007/b94608_15.
https://link.springer.com/content/pdf/10.1007/978-0-387-84858-7_15.pdf

[5] Marco Pota, Massimo Esposito, Giuseppe De Pietro, Designing rule-based fuzzy systems for classification in medicine, *Knowledge-Based Systems, Volume 124, Pages 105-132,* 2017, https://doi.org/10.1016/j.knosys.2017.03.006.