Applications of Support Vector Machine: Bais Reduction

Paola Quevedo

Computer Science Program, Roanoke College

1 Abstract

This study focuses on bias reduction when performing classification tasks on large datasets. Currently, there are state-of-the-art methods that provide positive results for this purpose such as Convolutional Neural Networks (Alvi, Zisserman, & Nellåker, 2018). However, the complexity of CNN models and their execution time has been part of the motivation to analyze other alternatives. This study is based on the implementation of Support Vector Machine algorithms for the classification of 50,884 images. The proposed method handles a dataset that contains biases in gender and age. One known method to reduce biases in Machine Learning applications is to increase the number of dimensions, with this in mind I created a Neural Network and explored the possibility of adding a third attribute to describe hair length. In addition, my approach to bias reduction involves training a Support Vector Machine model on a balanced dataset, and an analysis of the quality of the predictions. The same procedures were used on the unbalanced dataset for a 60/40 and 70/30 male to female ratio. I compared accuracy results using both the standard accuracy measures and Matthew's correlation coefficient (MCC) to analyze and determine the correct performance metric for the current problem. I hypothesized that my approach will provide a significant increase in the accuracy of the prediction using SVM. This research aims at making Machine Learning algorithms more transparent in the way that biased datasets are identified, and how the predictions become more reliable when the correct accuracy metrics are used.

2 Introduction

As machine learning is increasingly used in applications, machine learning algorithms are being studied more closely. Like many significant scientific advancements, the use of ML techniques has raised many significant ethical challenges. Ensuring that machine learning algorithms produce accurate classifications and predictions is vital as these automated processes can have a significant negative impact on a person's life. For example, ML systems have shown the capability to inherit racial and gender biases. In addition, ML systems have been used to predict, and often reveal, a user's attributes or even target their beliefs and psychological traits (Saltz et al., 2019). Therefore, working toward reducing biases in machine learning applications is an area of interest among the Computational and Data Science community.

3 Biases

Data bias in Machine Learning is considered as the type of error in which certain elements of a dataset are weighted or represented more heavily than others due to incorrect assumptions in the ML process. Biased datasets do not accurately represent the use case of the model, resulting in skewed outcomes and low accuracy levels (G. Harris, 2020).

4 Dataset

This research was conducted using a large publicly available dataset from the IMDb website which contains images of celebrities (Alvi et al., 2018). This is an unbalanced dataset with an estimated distribution of 60/40 with more males than females. The dataset was split into training, validation, and test sets with a random 80/10/10 split respectively. In addition to an overall skewed data with the gender attribute, the age groups are also unbalanced and the dataset presents biases toward younger women and older men.

5 Related Work

Research paper *Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings* was done using the same image dataset from the IMDb website that is used in this paper. They address

the use of large image datasets and how they are known to contain biases that cause models to generalize poorly to new, unseen data. The methods presented in this literature are based on a domain and task adaptation approach that would ignore a known bias in the dataset and improve the classification performance. They created two classifiers, one for age and one for gender. They trained two networks to first train solely on age data, then to train on age data, while removing gender-specific information from the network. They evaluated both networks on the unbiased test dataset and compared accuracies for age classification and prediction distributions for both genders (Alvi et al., 2018). The researchers introduce a "joint learning and unlearning (JLU) algorithm to learn a primary classification task, while simultaneously unlearning multiple spurious variations" (Alvi et al., 2018). They presented a method that uses a convolutional network (CNN) architecture where they used a "variation classification loss and a confusion loss which act in opposition to learn the classifier on the feature embedding and change the feature embedding to confuse the classifier, respectively" (Alvi et al., 2018). To compute probability measures they used functions such as softmax to predict the label of the image and entropy to calculate the confusion loss and overall loss which measure the distance from target to predicted values. It is stated that "cross-entropy was compared between the output distribution of classifiers that are trained to predict spurious variations and uniform distribution (Bourez, n.d.).

Their methodology involves working with feature representation which is the choice of features to represent patterns. This technique requires a task to learn a labeling pattern. The label assigned to a number depends upon the logical as well as the second and sixth bits of the binary representation of each number (Krishan, 2020).

5.1 Other functions

In the Turning a Blind Eye literature, they used functions such as softmax to predict the label of the image and entropy to calculate the confusion loss and overall loss which measure the distance from target to predicted values. It is stated that "cross-entropy was compared between the output distribution of classifiers that are trained to predict spurious variations and uniform distribution (Bourez, n.d.).

5.2 Their findings

They were successful at demonstrating that the feature representation of the baseline network that was trained to classify age is separable by gender, showing that the bias was learned in the training data. Later, gender was unlearned, and it was proven that the feature representation was no longer separable by gender and therefore demonstrating that the bias was removed. The methods implemented helped in significantly improving (by up to 20% classification accuracies. By removing each known bias from the feature representation of the network, it will allow networks trained on biased data to generalize better to unbiased settings (Alvi et al., 2018).

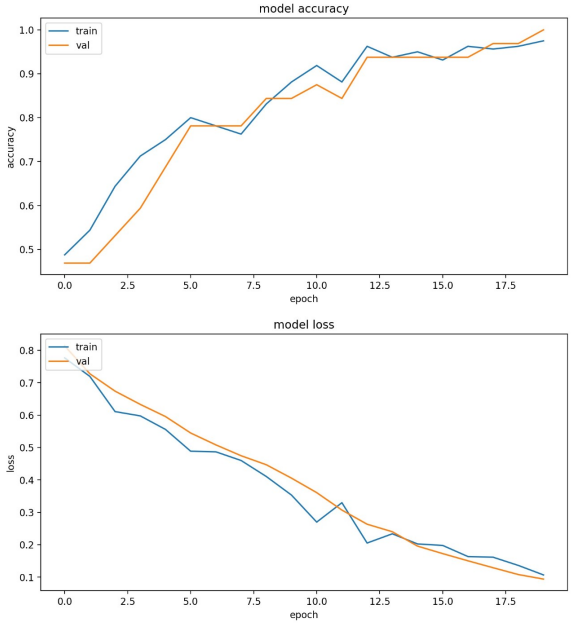
6 Neural Network and SVM Experiments

In an effort to prove that my proposed method could potentially decrease biased predictions from this data set using a different approach than the one used in the mentioned literature, I experimented with a neural network to add an extra attribute to the dataset that would describe hair in the images. This neural network was successfully adjusted and trained to predict the new tag using a transfer learning process that requires the use of a pre-trained model called ResNet-512 belonging to the Keras API and it was introduced by the paper *Deep Residual Learning for Image Recognition*. Several modules and functions were involved to ensure that the data is pre-processed and that the NN layers are treated properly. In addition, I created a Suppor Vector Machine model for classifying images with gender and age attributes. The process of creating an SVM model involved preparing the dataset as well as trying different parameters for the optimization stage to produce the best possible classifier. I evaluated the accuracy of the predictions by comparing both the Standard and Mathew's Correlation Coefficient scores to determine the metrics that would provide the most reliable representation of the distribution of classes and errors in the dataset.

6.1 Neural Network With Hair

ResNet is a pre-trained model with pre-trained weights that were trained on a large database named ImageNet (He, Zhang, Ren, & Sun, 2015). This network handles classification tasks with 1,000 classes and it was adjusted to work with only two classes. The implementation was done using Tensorflow (*Tensorflow*, n.d.), and Keras transfer learning and fine-tuning techniques

(Team, n.d.). To train this model for the current problem, a balanced set of 100 images were used. Data augmentation was performed by applying different rotation, zoom and mirror effects. Once the images were sent to the preprocessing function of ResNet, the images were then sent to the network. Convergence was noticed after 20 iterations. This is evidence that learning was successful.



The large dataset of files was later labeled with the hair attribute to test its functionality, however, it will not be used to study biases in this paper as previously mentioned. It mainly serves as an option to add more labels to the dataset if it is needed.

6.2 Conclusion for the Neural Network With Hair Experiment

The intention for adding an extra attribute to the data, was to evaluate whether the Support Vector Machine model could be used for more complex datasets. However, after considering the possible biases that could be introduced to the dataset due to the naturally high percentage of women who have long hair compared to men, I determined that a hair length attribute would not add any insight or useful information when predicting gender.

The goal for machine learning applications is for the predictions to reflect the true population in the context being studied. The expected outcome with the additional label is that it would cause more biases in the dataset. As it is discussed by Researches at the National Institute of Standards and Technology (NIST), they recommend widening the scope of where we look for the source of biases — beyond the machine learning processes and data used to train AI software to the broader societal factors that influence how technology is developed" (Sarah.henderson@nist.gov, 2022). They emphasize how bias manifests itself not only in AI algorithms and the data used to train them but also in the societal context in which AI systems are used. Hair length is therefore not a proper attribute to consider for my research problem since it does not have the potential to reduce the biases in the dataset and the context in which it is being studied.

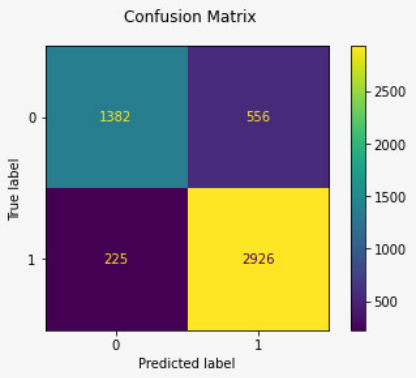
6.3 Support Vector Machines (SVM)

Support Vector Machines are a set of supervised learning methods that are used to perform common tasks in machine learning such as classification, regression, and outliers detection. SVMs are different from other classification algorithms because of the way they choose the decision boundary that maximizes the distance from the nearest data points of all the classes. The decision boundary created by SVMs is called the maximum margin classifier or the maximum margin hyperplane (McGregor, 2020). In this paper, Kernel SVM is used since it has more flexibility for non-linear data and I will use Gaussian Radial Basis Function (RBF) for the kernel represented by equation:

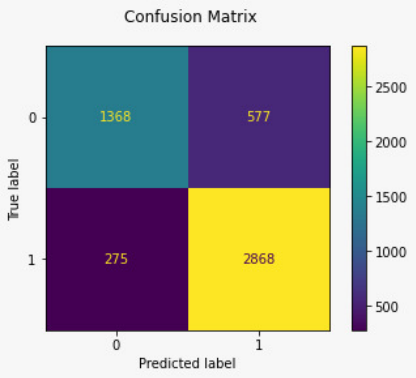
$$f(X1, X2) = exp(-gamma * ||X1 - X2||^2)$$

6.3.1 SVM Model

I experimented with the unbalanced dataset, the SVM model was trained and it produced predictions where more females were wrongly classified as males, this is because of the higher number of males in the dataset. The model learns to predict based on the highest probable case which will follow a similar proportion to the existing biases. Females are represented by 0 and males by 1.

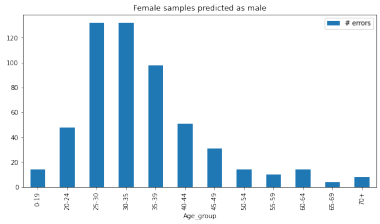


((a)) Validation

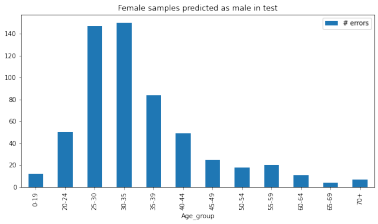


((b)) Test

Figure 1: Confusion Matrix for unbalanced sets

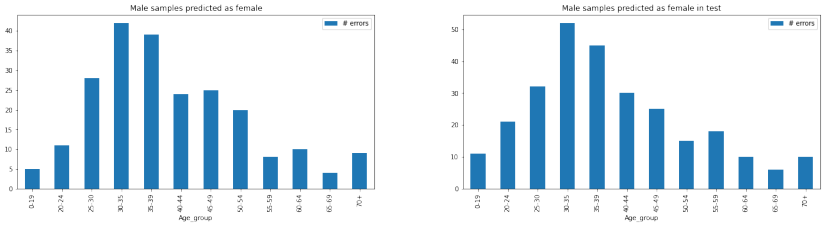


((a)) Validation



((b)) Test

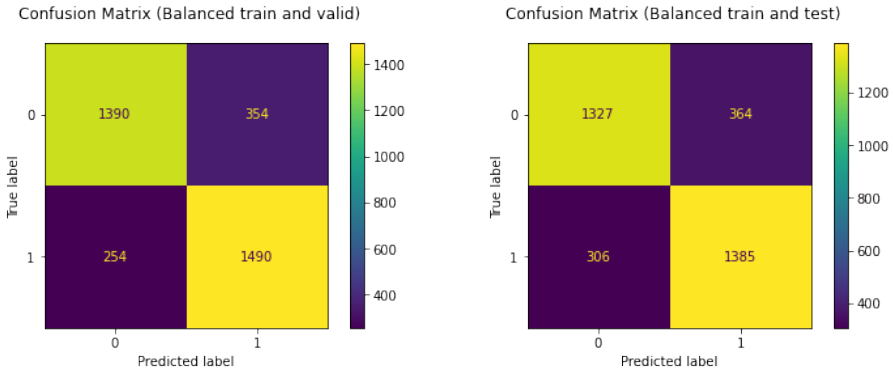
Figure 2: Distribution of females predicted as males



((a)) Validation ((b)) Test

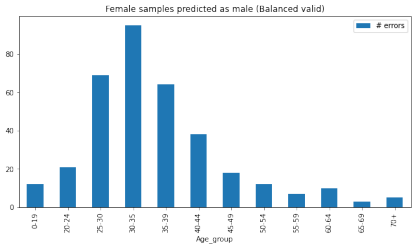
Figure 3: Distribution of males predicted as females

I created a function to balance the distribution for gender to reflect a 50/50 gender representation. The dataset was then reduced to 27,294 images for training. After training, more errors occurred possibly due to having fewer data. However, the errors showed a distribution close to 50/50. The standard accuracy measure was used.

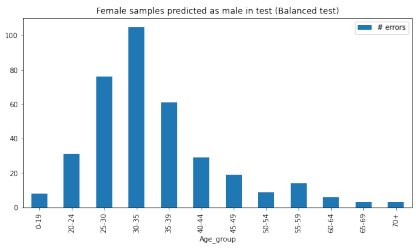


((a)) Validation ((b)) Test

Figure 4: Confusion Matrix for balanced sets

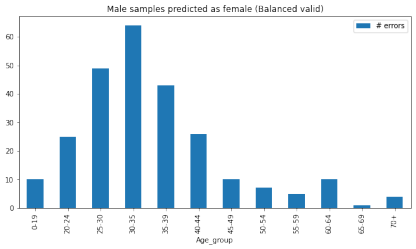


((a)) Validation

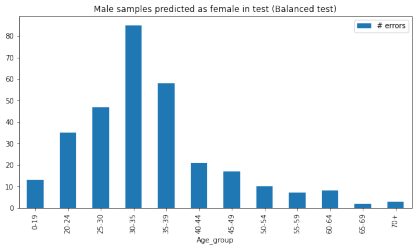


((b)) Test

Figure 5: Distribution of females predicted as males



((a)) Validation



((b)) Test

Figure 6: Distribution of males predicted as females

In this experiment I do a comparison of accuracy metrics with a 50/50 and 60/40 distribution to analyze the level of accuracy in both cases. Matthew's correlation coefficient (MCC) is being considered as it is a measure of the quality of binary and multi-class classifications, and in some cases, it has shown to be more reliable than other metrics when the dataset is unbalanced (Zhu, 2020).

Standard Measures For The Test Set

Balanced dataset Metrics			
Class	F1 score	Recall	Presicion
Female	76%	78%	73%
Male	84%	82%	86%

Unbalanced dataset Metrics			
Class	F1 score	Recall	Presicion
Female	80%	78%	81%
Male	81%	82%	79%

Mathew's Correlation Coefficient	
Balanced (50/50)	0.60
Unbalanced (60/40)	0.64

Age_group	# errors
0-19	12
20-24	50
25-30	147
30-35	150
35-39	84
40-44	49
45-49	25
50-54	18
55-59	20
60-64	11
65-69	4
70+	7

((a)) Females predicted as males - Unbalanced

Age_group	# errors
0-19	11
20-24	21
25-30	32
30-35	52
35-39	45
40-44	30
45-49	25
50-54	15
55-59	18
60-64	10
65-69	6
70+	10

((b)) Males predicted as females - Unbalanced

Age_group	# errors
0-19	8
20-24	38
25-30	118
30-35	113
35-39	61
40-44	29
45-49	19
50-54	9
55-59	14
60-64	6
65-69	3
70+	3

((c)) Males predicted as females - Balanced

Age_group	# errors
0-19	13
20-24	35
25-30	47
30-35	85
35-39	84
40-44	64
45-49	56
50-54	41
55-59	45
60-64	33
65-69	23
70+	28

((d)) Males predicted as females - Balanced

Figure 7: Unbalanced Vs. Balanced

7 Conclusion for SVM Experiment

In this research it was confirmed that there are biases in the overall gender distribution with approximately 60/40 males and females respectively as well as an imbalance in samples within the age groups. Although this might not be considered significantly unbalanced in some scenarios, it is crucial that machine learning algorithms provide the most accurate results when making predictions. I would expect to see more errors as the dataset distribution becomes even more unbalanced. It is of everyone's interest that machine learning algorithms provide the most reliable results to allow for additional steps to be considered in the event that a concerning error percentage is reported. After evaluating the predictions from the unbalanced dataset there is a higher number of errors in the predictions where the algorithm predicted females as males. This is due to the fact that there are more males in the dataset. However, when looking at the errors of males wrongfully predicted as females, more errors occurred among ages ranging from 25-35 which leads me to believe that the large number of females in

those young age groups is causing biased predictions for the images as well. When analysing the errors from the balanced dataset, more errors occur possibly due to the reduced number of samples. However, the errors are better distributed. When using the standard measure of accuracy the percentages were very similar in both distributions. This can serve as evidence that the biases in the predictions are not properly reflected in the results when using the standard measure. When evaluating the accuracy with Mathew's Correlation Coefficient it is showing more reliable report. The accuracy score in this case was 0.60 for balanced and 0.64 for the unbalanced set. The Matthews correlation coefficient (MCC), is a more reliable statistical rate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset compared to the f1 score which ignores the True Negatives in the calculation (Chicco & Jurman, 2020). My SVM models did not prove to be more reliable for predictions on this dataset of images than the convolutional nueral network used by the previous researchers. It did not prove to be time efficient when trying to train this large dataset multiple times to adjust the parameters. But, I do believe there is opportunity to continue to explore this models on the image dataset. One possible approach to reducing the biased results could be to continue to work with the hyper-parameters like C or Gamma which control the movement of the SVM decision boundary.

References

Alvi, M. S., Zisserman, A., & Nellåker, C. (2018). Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. *CoRR*, *abs/1809.02169*. Retrieved from http://arxiv.org/abs/1809.02169

Bourez, C. (n.d.). *Why targets 0 and 1 in machine learning ?* Retrieved from http://christopher5106.github.io/deep/learning/2018/10/20/course-zero-deep-learning.html

Chicco, D., & Jurman, G. (2020, Jan). *The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation - bmc genomics.* BioMed Central. Retrieved from https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6413-7

G. Harris, C. (2020). Mitigating cognitive biases in machine learning algorithms for decision making. In *Companion proceedings of the web conference 2020* (p. 775–781). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3366424.3383562

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition.* arXiv. Retrieved from https://arxiv.org/abs/1512.03385 doi: 10.48550/ARXIV.1512.03385

Krishan. (2020, Mar). *Feature representation.* Retrieved from https://iksinc.online/tag/feature-representation/

McGregor, M. (2020, Jul). *Svm machine learning tutorial – what is the support vector machine algorithm.* freeCodeCamp.org. Retrieved from https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/

Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., ... Beard, N. (2019, aug). Integrating ethics within machine learning courses. *ACM Trans. Comput. Educ.*, *19*(4). Retrieved from https://doi.org/10.1145/3341164 doi: 10.1145/3341164

Sarah.henderson@nist.gov. (2022, Mar). *There's more to ai bias than biased data, nist report highlights.* Retrieved from https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights

Team, K. (n.d.). *Keras documentation: Transfer learning amp; fine-tuning.* Retrieved from https://keras.io/guides/transfer$_l$*earning*/

Tensorflow. (n.d.). Retrieved from https://www.tensorflow.org/

Zhu, Q. (2020). On the performance of matthews correlation coefficient (mcc) for imbalanced dataset. *Pattern Recognition Letters*, *136*, 71-80. Retrieved from https://www.sciencedirect.com/science/article/pii/S016786552030115X doi: https://doi.org/10.1016/j.patrec.2020.03.030