COMP30027 Report

Anonymous

1. Introduction

This report will explore and analyse different supervised Machine Learning methods to predict the ratings of books through the given training and testing datasets.

Classifiers to be discussed are Linear Support Vector Machine (SVM) and 0R Baseline Classifier as the base model. This report will analyse the performance of each classifier by constructing, evaluating, and error analysing the classifiers.

2. Data Description

The training and testing datasets are provided and extracted from Goodreads.

2.1 Data Overview

The training and testing datasets contain 23063 and 5766 instances, respectively. The datasets provided contain the book features such as name, authors, publish year, publish month, publish day, publisher, language, page numbers, and description, with rating labels only present in the training datasets.

2.1.1 Text Features

We are also given the preprocessed text features obtained from different text encoding methods (Count Vectorizer and Doc2Vec) for name, authors, and description for both training and testing datasets.

3. Methodology

Text features used are the one encoded using Count Vectorizer, which provides sparse matrices of name, authors, and description features.

3.1 Preprocessing

As the datasets contain missing values in both publisher and language features, we assign a constant value to handle the missing value. For publisher, we replace the missing values with 'Unknown', while for language, we fill in 'Missing' to replace the missing values. Also, the text-type features such as language and publisher are converted to sparse matrices using the Term Frequency – Inverse Document Frequency (TF-IDF) vectorizer. Then, the numeric features such as publish year, publish month, publish day, and page number are combined with the sparse matrices of name, authors, description, publisher, and language.

3.2 Feature Selection

While selecting the feature, several things that are taken into consideration. Firstly, the impact of missing values on the performance of the model. This is because language attribute has so many missing values (17202 data), and publisher attribute has in total of 148 missing data. Secondly, how strong these attributes are in predicting the class label. Therefore, we run our model on four different types of features: data with all features including language and publisher, all features excluding publisher, all features excluding language, and all features excluding both publisher and language.

3.3 Model Selection

Machine learning methods to be used are 0R Baseline Classifier and Linear SVM. 0R Baseline Classifier is popular as it is a straightforward baseline. It predicts the most common label for all instances in the datasets. Moreover, it is very effective to be used to assess more complex models. On the other hand, Linear SVM is known for its effectiveness in handling high-dimensional and large datasets. In addition, Linear SVM is also robust to outliers as it uses margin to optimise the decision boundary. It uses the hyperparameter C to control the margin and allow a few misclassifications in the training data to achieve better performance.

3.4 Results

As 0R Baseline Classifier does not consider any features and only determines the majority class label on the datasets, we use Linear SVM to decide what features to use. We tried different hyperparameters C, and the performance in accuracies is shown in Figure 1.

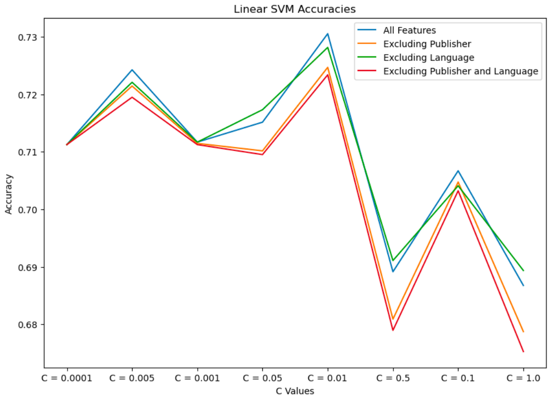

Figure 1 – Linear SVM Accuracies with different features and values of C.

According to the model used, we then decided to use all features and tune the hyperparameter C to 0.01 as it has higher accuracy among other features and C values. In summary, both publisher and language attributes have impacts on predicting the class label as the accuracy increases when both features are included. Thus, we conclude that both features are essential for the model despite their missing values.

4. Model Training and Evaluation

4.1 0R Baseline Classifier

4.1.1 Training Process

The original train datasets are split into new train and test sets using the train test split function with train size = 0.8, test size = 0.2, and random state = 42.

4.1.2 Implementation

We fit the model to the new train sets and apply it to the new test sets.

4.1.3 Evaluation Metrics

The accuracy obtained from the accuracy score function is 0.7113, with further evaluation metrics shown in Table 1.

	Precision	Recall	F1-Score
Macro Average	0.24	0.33	0.28
Weighted Average	0.51	0.71	0.59

Table 1 – Evaluation Metrics of 0R Classifier.

Aligned with the principle of this model, the accuracy, precision, and recall are determined by the proportions of training data that belong to the majority class label.

4.1.4 Results and Limitations

While the accuracy of 0R baseline Classifier is reasonably high, it is important to note that this Classifier only predicts the majority label of the instances without acknowledging the importance and relationship of each feature.

4.2 Linear SVM

4.2.1 Training Process

The combined matrix datasets are split into new sets of train and test sets using the train test split function with train size = 0.8, test size = 0.2, and random state = 42. Additionally, MaxAbsScaler is used to scale the new train and test set as this scaler allows us to preserve the sparsity structure and reduce the influence of outliers. Then, we fit the model with a range of C values (from 0.00001 to 1.0).

4.2.2 Evaluation Metrics

From the tuning of C, we then use C = 0.01 to fit the test set as it has better performance in accuracy (Figure 2).

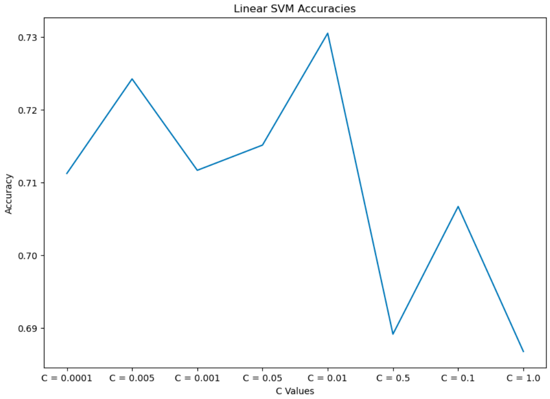

Figure 2 - Linear SVM with different values of C.

4.2.3 Results

Compared to 0R Baseline Classifier, Linear SVM with C = 0.01 perform better with accuracy = 0.7305 (Further evaluation metrics are provided in Table 2). The overall performance for Linear SVM is better than the baseline model. In conclusion, this model shows relatively high improvement compared to the simplest approach, 0R Baseline Classifier.

	Precision	Recall	F1-Score
Macro	0.79	0.40	0.41

Average			
Weighted Average	0.72	0.73	0.67

Table 2 – Evaluation Metrics of Linear Support Vector Machine (C= 0.01).

5. Discussion

5.1 Model Interpretation

Linear SVM algorithm uses hyperplane as a decision boundary that categorizes data points into different class labels. It also uses margin as the distance between the decision boundaries. This margin allows some misclassification of the data points to correctly classify more points. This model is good and effective in high-dimensional data, including our book datasets. Additionally, it has a strong theoretical foundation and works well with linearly separable data.

5.2 Limitation

However, this model is not limited to any drawbacks. Firstly, our model assumes that the relationship between the features and class label is linear, which may lead to bias and inaccuracy if the datasets actually have complex non-linear patterns. Secondly, our model is also not limited to overfitting, which can affect the performance of the model.

5.3 Error Analysis

Misclassification is then identified to understand the models further.

For 0R Baseline Classifier, it is expected that the model will predict all the class labels as the most frequent class (4.0 in this case) as 0R Model do not consider any relationship or pattern from the features.

However, for Linear SVM, after categorizing each feature, it is found that the model makes the most error in publish day = 1 (540 error data), publish month = 1 (137 error data), publish year = 2006 (141 error data), page number = 256 (50 error data), language = 'Missing' (1002 error data).

It can be seen that the missing value from language has the most impact on the error made. However, as stated above, the accuracy of prediction is higher than those if we do not use the language feature.

Hence, one way to address this error is that more data on language can be collected to strengthen the learning of the model, which will improve the performance of the model.

6. Conclusions

6.1 Summary

In summary, Linear SVM is a good supervised Machine Learning method for classifying and predicting class labels. Linear SVM learns the features of datasets through vectors and uses decision boundary in the feature space to assign each data point to each class label correctly. Compared to 0R Baseline Classifier, Linear SVM succeeded in providing more insightful information. It captures the relationship between features and target variables well and is able to achieve a relatively good performance.

6.2 Future Work

While this project has provided information on the performance of the two models discussed, this study is very open for future development and improvement to address any limitations on this study.

Future studies may collect more varied datasets with more class labels to examine the model in more depth. In addition, the relationship between each feature and each importance in predicting the class label may also be captured to confidently assess the pattern. Moreover, future investigations should explore additional features that contribute to a better understanding of the datasets. Furthermore, in-depth feature selection might be applied to future studies to correctly examine the importance and impact of each feature. Finally, while our model produces a very good result, it is always good to try other comparative algorithms models to assess the supervised Machine Learning method in many other aspects.

7. References

Goodreads. (n.d.). *Meet your next favorite book*. Goodreads. https://www.goodreads.com/