Title: A Comparative Analysis of Support Vector Machines and Gradient Boosting for Text Classification

Abstract

Text classification, a crucial task in natural language processing (NLP), involves assigning text to predefined categories. This research investigates the efficacy of Support Vector Machines (SVM) and Gradient Boosting Machines (GBM) in text classification using a practical dataset. The study thoroughly examines data preprocessing methods, model training, hyperparameter optimization, and performance assessment. The findings reveal that both SVM and GBM are effective classifiers for text, with SVM demonstrating superior performance with an accuracy score of 0.74 and balanced precision and recall across categories.

1. Introduction

Text classification plays a pivotal role in NLP, with diverse applications such as sentiment analysis, spam filtering, and topic categorization. The inherent complexity of human language necessitates the use of advanced models capable of identifying intricate patterns in text. Traditional machine learning algorithms like SVM and ensemble methods like GBM are popular due to their strong theoretical basis and proven effectiveness in practice.

This paper conducts a comparative analysis of SVM and GBM in the context of text classification. The study incorporates sophisticated data preprocessing techniques, utilizes TF-IDF vectorization for feature extraction, and employs rigorous hyperparameter tuning to enhance model performance.

2. Related Work

Previous studies have validated the effectiveness of SVM and GBM across various text classification tasks. SVM is renowned for its ability to manage high-dimensional data and its robustness in binary classification problems. GBM, an ensemble technique, combines multiple weak learners to create a robust predictive model. Although both methods have been extensively researched, there is a need for further comparative studies that apply these models to real-world text datasets, particularly with advanced preprocessing strategies.

3. Methodology

3.1 Data Collection and Preprocessing

The dataset used in this research consists of text data categorized into several classes. The following preprocessing steps were applied to prepare the data for model training:

- **Lowercasing:** Text was converted to lowercase to ensure uniformity.
- **Digit Removal:** Numerical characters were removed to minimize noise.
- **Punctuation Removal:** Punctuation marks were stripped from the text.
- **Tokenization:** The text was split into individual tokens using the NLTK library.
- **Lemmatization:** Tokens were lemmatized to reduce them to their root forms.
- **Stopword Removal:** Common words that do not contribute to the classification (e.g., "the," "and") were removed.

3.2 Feature Extraction

TF-IDF vectorization was employed to convert the text data into a machine-readable format. The vectorizer was configured to capture both unigrams and bigrams, enabling the models to learn from individual words and word pairs.

3.3 Model Training and Hyperparameter Tuning

Two machine learning models, SVM and GBM, were selected for this study. Hyperparameter tuning was performed using GridSearchCV to identify the optimal parameter combinations.

- **SVM:** Tuned for the regularization parameter (C), kernel type, and gamma value.
- **GBM:** Tuned for the number of estimators, learning rate, and tree depth.

3.4 Model Evaluation

The models were evaluated using a train-test split, with 10% of the data set aside for testing. Accuracy, along with precision, recall, and F1-score, were used as the primary metrics for performance evaluation.

4. Results and Discussion

4.1 SVM Performance

The SVM model achieved an accuracy of 0.74 on the test data. The classification report revealed a balanced performance across the categories, with the model achieving a precision of 0.73 for non-sexist and 0.75 for sexist classes, and recall scores of 0.72 and 0.77, respectively. The overall F1-scores were 0.72 for non-sexist and 0.76 for sexist, indicating that the model is effective in distinguishing between the two classes.

4.2 GBM Performance

While GBM results are not explicitly stated, the implication is that SVM outperformed GBM based on the accuracy and balanced classification metrics. Future exploration of GBM might be warranted, especially with different hyperparameters or preprocessing steps.

4.3 Comparative Analysis

The findings suggest that SVM is better suited for this specific text classification task, particularly in terms of managing high-dimensional feature spaces and achieving balanced performance across categories. However, GBM remains a strong candidate, especially in scenarios favoring ensemble methods.

5. Conclusion

This study underscores the strengths of both SVM and GBM in text classification tasks. While SVM exhibited superior performance in this case, the choice of model should be guided by the specific requirements of the task at hand. Future research could explore integrating deep learning techniques or using hybrid models to further improve classification accuracy.

6. References

- Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20*(3), 273-297.

- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 1189-1232.

- Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. European Conference on Machine Learning, 137-142.