Report: Feature Selection by Integrating Document Frequency with Genetic Algorithm for Amharic News Document Classification

Source

Title: *"Feature selection by integrating document frequency with genetic algorithm for Amharic news document classification"*
Journal: PeerJ Computer Science (2022)
DOI: https://doi.org/10.7717/peerj-cs.961
Authors: Demeke Endalie, Getamesay Haile, Wondmagegn Taye Abebe

Executive Summary

This research introduces a **hybrid feature selection method** combining **Document Frequency (DF)** and **Genetic Algorithm (GA)** for **Amharic text classification**. The study addresses the challenge of high-dimensional text data in low-resource languages like Amharic and demonstrates that the proposed **DFGA method** outperforms existing filter-based and hybrid methods in classification accuracy while significantly reducing the number of features.

Key Points from the Report

1. Problem & Motivation

- **Amharic is a low-resource language** with limited NLP tools and datasets.
- **Text classification** faces the **curse of dimensionality**, requiring effective feature selection.
- **Existing methods** (filter-based) do not consider classifier performance and require manual threshold tuning.

Report: Feature Selection by Integrating Document Frequency with Genetic Algorithm for Amharic News Document Classification

- **Goal:** Develop a hybrid FS method that improves accuracy and reduces feature count.

2. Proposed Method: DFGA (Document Frequency + Genetic Algorithm)

- **Step 1:** Use **Document Frequency (DF)** as a **filter method** to select an initial set of relevant features.
- **Step 2:** Apply **Genetic Algorithm (GA)** as a **wrapper method** to find the optimal feature subset based on classifier performance.
- **Advantages:**
  - Combines speed of filter methods with accuracy of wrapper methods.
  - Considers classifier feedback during feature selection.
  - Reduces dimensionality effectively.

3. Experimental Setup

- **Dataset:** 13 categories of Amharic news from Ethiopian News Agency (3,158 documents).
- **Preprocessing:** Normalization, stop-word removal, stemming (using HornMorpho).
- **Representation:** Bag-of-Words (BoW) with TF-IDF weighting.
- **Classifier:** Extra Tree Classifier (ETC) – also tested with Random Forest and Gradient Boosting.
- **Comparison Methods:** DF, IG, CHI, PCA, GA, and hybrid combinations (IG+CHI+DF, IG+CHI+DF+PCA).

4. Key Results

- **Best Accuracy: 89.68%** with DFGA + ETC (80/20 train-test split).
- **Outperforms:**
  - DFGA vs. IG+CHI+DF+PCA: **+1.01%**
  - DFGA vs. GA alone: **+2.47%**
  - DFGA vs. IG+CHI+DF: **+3.86%**
- **Feature Reduction:** DFGA selected only **100 features**, far fewer than other methods (e.g., PCA: 1,226, GA: 230).

Report: Feature Selection by Integrating Document Frequency with Genetic Algorithm for Amharic News Document Classification

- **Best Classifier:** ETC outperformed RFC and GBC with DFGA.

5. Contributions

1. **Novel Hybrid FS Method:** First to combine DF and GA for Amharic text classification.
2. **Improved Accuracy & Efficiency:** Higher accuracy with fewer features.
3. **Resource-Friendly:** Suitable for low-resource language processing.

6. Challenges & Insights

- **Filter methods alone** ignore feature interactions and classifier impact.
- **Wrapper methods (GA)** are computationally heavy but more accurate.
- **Hybrid approach** balances speed and performance.

7. Future Work

- Test DFGA on larger datasets and more categories.
- Apply to other low-resource languages.
- Explore integration with deep learning models.

Conclusion

The **DFGA method** effectively combines the simplicity of DF with the optimization power of GA, resulting in **higher classification accuracy and significant feature reduction** for Amharic news classification. This approach is promising for NLP applications in low-resource languages and can enhance tasks like document organization, topic extraction, and information retrieval.