Optimization of Feature Selection Using Genetic Algorithm in Naive Bayes Classification for Incomplete Data

Authors:

Bain Khusnul Khotimah, Miswanto Miswanto, Herry Suprajitno

Publication:

International Journal of Intelligent Engineering & Systems, Vol.13, No.1, 2020

Link: https://doi.org/10.22266/ijies2020.0229.31

Objective:

To improve classification accuracy for high-dimensional datasets with missing values by combining **Self-Organizing Map Imputation (SOMI)** with a **Genetic Algorithm (GA)** for feature selection and **Naive Bayes Classification (NBC)**.

Key Contributions:

1. **Hybrid Approach (SOMI-GANB):**
    a. Uses SOMI to impute missing values by clustering and replacing them with prototype weights.
    b. Applies GA to select optimal feature subsets.
    c. Uses Naive Bayes for classification.
2. **Methodology:**
    a. **Preprocessing:** SOMI handles missing data for mixed (continuous and discrete) datasets.
    b. **Feature Selection:** GA selects relevant features using a fitness function based on classification accuracy.
    c. **Classification:** Naive Bayes classifies the imputed and reduced dataset.

Optimization of Feature Selection Using Genetic Algorithm in Naive Bayes Classification for Incomplete Data

3. **Experiments:**
    a. Tested on 5 datasets (Glass, Heart Disease, WDBC, Sonar, Record Linkage Comparison Patterns) with missing rates from 5% to 40%.
    b. Compared against other imputation methods: Mean, Modus, Median, Hot Deck, and methods without feature selection.
4. **Results:**
    a. SOMI-GANB achieved **up to 90%+ classification accuracy**, outperforming other methods.
    b. SOMI reduced imputation error by up to 10% compared to traditional methods.
    c. Feature selection with GA improved accuracy and reduced dimensionality without significant information loss.
5. **Conclusions:**
    a. The hybrid SOMI-GANB method is effective for incomplete, high-dimensional, and mixed datasets.
    b. GA optimizes feature selection, enhancing Naive Bayes performance.
    c. Future work includes optimizing feature weighting and hybridizing GA with other evolutionary algorithms.