Classification of Breast Cancer Data Using Machine Learning Algorithms

Overview of the Study

The study's objective is to assess the classification performance of machine learning algorithms applied to medical datasets, specifically focusing on breast cancer classification. Machine learning offers computational strategies that can yield optimal solutions without explicit programming, and its use in medical data classification has been on the rise.

Dataset and Methodology

The dataset used is the Breast Cancer Wisconsin (Original) dataset, consisting of 699 instances and 10 attributes. The classification is conducted using the Weka software, which provides tools for data preparation, classification, clustering, and other machine learning tasks.

The algorithms investigated include:

- **k-Nearest Neighbors (k-NN)** with varying k-values
- **Naïve Bayes (NB)** using default settings in Weka
- **Support Vector Machine (SVM)** using different cost (C) and gamma (ɣ) values and kernel functions.

The classifiers are evaluated using 10-fold cross-validation to prevent overfitting and to ensure accurate classification of new datasets.

Results

1. **k-Nearest Neighbors (k-NN)**:
    - The classifier with a k-value of 3 achieved the highest accuracy of 96.85%. It was able to correctly classify 445 out of 458 non-cancerous cases and 232 out of 241 cancerous cases.
2. **Naïve Bayes (NB)**:
    - The NB classifier achieved an accuracy of 95.99%, with fewer tunable parameters compared to other algorithms.
3. **Support Vector Machine (SVM)**:
    - The C-SVM classifier with Radial Basis Kernel (ɣ = 2^-15) and cost parameter C = 2^15 achieved the same accuracy as k-NN with 96.85%.
    - Other SVM variants produced lower accuracies, highlighting the importance of selecting optimal parameters for SVM.

Conclusion

The study concludes that both k-NN with k = 3 and C-SVM with specific cost and gamma values are highly effective in classifying breast cancer data, achieving an accuracy of 96.85%. Future studies could explore these algorithms using different programming environments such as Python or R, which might offer performance improvements over Java-based Weka.

This research contributes to the ongoing efforts in applying machine learning techniques for medical data classification, particularly for breast cancer diagnosis, and underscores the potential of machine learning to improve the accuracy and efficiency of medical diagnoses.

On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset

Overview of the Study

The goal of the study is to compare the effectiveness of different machine learning algorithms in classifying breast cancer cases as benign or malignant. The dataset used for this study is widely employed in breast cancer research, consisting of 569 instances, of which 212 are malignant and 357 are benign.

Dataset and Methodology

The machine learning models explored in this study include:

- **GRU-SVM**: A hybrid of a Gated Recurrent Unit (GRU) and Support Vector Machine (SVM)
- **Linear Regression**
- **Multilayer Perceptron (MLP)**
- **Nearest Neighbor Search**
- **Softmax Regression**
- **Support Vector Machine (SVM)**

The dataset was split into 70% for training and 30% for testing. The features of the dataset were standardized before being used in the algorithms. Stochastic Gradient Descent (SGD) was used to train most of the algorithms, except for GRU-SVM, Nearest Neighbor, and SVM.

Results

1. **Multilayer Perceptron (MLP)**:
    - Achieved the highest accuracy of 99.04%.
2. **Softmax Regression**:
    - Performed well with an accuracy of 97.66%.
3. **L2-SVM**:
    - Achieved an accuracy of 96.09%, which was better than previous studies using SVM with a Gaussian Radial Basis Function (RBF).
4. **GRU-SVM**:
    - Achieved an accuracy of 93.75%, lower than the others due to potential difficulties in generalizing linear data.
5. **Nearest Neighbor Search (L2)**:
    - Achieved an accuracy of 94.74%.

Conclusion

The study concludes that all machine learning algorithms tested in the study are effective for breast cancer classification, with MLP providing the highest accuracy. The results suggest the use of cross-validation techniques for further model improvements, along with optimal hyper-parameter tuning for better performance.

This study confirms that machine learning can significantly assist in the accurate diagnosis of breast cancer.