Kapil Yadav (B22AI024)

Arsewad Bhagwan (B22AI010)

Task 0: Classification Algorithms Evaluation for IAPRTC-12 Dataset

1. Introduction

This report presents the evaluation of six classification algorithms on the IAPRTC-12 dataset for multi-label image classification. The primary objective was to establish baseline performance metrics for these algorithms before investigating their robustness against adversarial attacks in subsequent tasks.
The following algorithms were evaluated:

- Linear SVM
- Logistic Regression
- Softmax Regression
- Decision Tree
- Weighted KNN
- An ensemble of all these algorithms

2. Dataset Overview

The IAPRTC-12 dataset contains images with text descriptions and multi-label annotations. For this evaluation, we used the provided feature representations rather than raw images.
Dataset Dimensions:

- Training set: 17,665 samples × 2,048 features, with 291 possible labels
- Testing set: 1,962 samples × 2,048 features, with 291 possible labels

This represents a challenging multi-label classification task due to the high feature dimensionality and large label space.

3. Methodology

3.1 Data Preprocessing

Two key preprocessing steps were implemented:
Feature Scaling: Standard scaling was applied to normalize features to zero mean and unit variance, which is particularly important for distance-based algorithms like KNN and models like SVM and logistic regression.

Dimensionality Reduction: PCA was applied to reduce the feature space from 2,048 to 200 dimensions while preserving 72% of the variance. This significantly reduced computational requirements while maintaining most of the information content.

3.2 Model Configuration

Each algorithm was implemented using scikit-learn with the following configurations:
Linear SVM: One-vs-Rest classification strategy for multi-label classification.
Logistic Regression: One-vs-Rest strategy with default parameters and increased maximum iterations.
Softmax Regression: Multinomial logistic regression with 'lbfgs' solver.
Decision Tree: Maximum depth of 5 to prevent overfitting, minimum samples split of 5, and minimum samples per leaf of 2.
Weighted KNN: Distance-weighted voting with 5 neighbors and Euclidean distance metric.
Ensemble: Majority voting across all individual models, requiring at least 3 models to predict a label for it to be included in the final prediction.

3.3 Evaluation Metrics

The following metrics were used to evaluate performance:

- Accuracy (exact match)
- Precision (micro and macro averaged)
- Recall (micro and macro averaged)
- F1 score (micro and macro averaged)
- Hamming Loss (fraction of incorrect labels)

4. Results and Analysis

4.1 Performance Metrics

Model	Accuracy	Precision (micro)	Precision (macro)	Recall (micro)	Recall (macro)	F1 (micro)	F1 (macro)	Hamming Loss
Linear SVM	0.0183	0.7566	0.3211	0.2255	0.1026	0.3475	0.1398	0.0164
Logistic Regression	0.0000	0.0704	0.0554	0.8624	0.8201	0.1302	0.0975	0.2230
Softmax Regression	0.0000	0.0704	0.0554	0.8624	0.8201	0.1302	0.0975	0.2230
Decision Tree	0.0102	0.5300	0.3606	0.2440	0.1702	0.3342	0.2157	0.0188

Weighted KNN	0.0607	0.6498	0.5554	0.3616	0.2669	0.4647	0.3362	0.0161
Ensemble	0.0291	0.5572	0.4492	0.4347	0.3169	0.4884	0.3476	0.0176

4.2 Analysis of Results

Accuracy: Weighted KNN achieved the highest accuracy (0.0607), followed by the Ensemble approach (0.0291). The low accuracy values across all models reflect the difficulty of exactly matching all 291 labels.
Precision vs. Recall Trade-off: Linear SVM showed the highest precision (micro: 0.7566) but lower recall, while Logistic and Softmax Regression demonstrated the opposite pattern with very high recall

5. References

- Scikit-Learn Documentation
- Linear SVM (LinearSVC) Guide
- Logistic Regression Guide
- Softmax Regression (Multinomial Logistic Regression)
- Decision Tree Classifier
- K-Nearest Neighbors (KNN) Classifier
- One-vs-Rest (OvR) Strategy
- Ensemble Methods in Scikit-Learn