CSCI4150U: Data Mining

Lab 03 - Classification and Clustering Evaluation

Syed Naqvi
Student ID: 100590852

November 1, 2024

Abstract

This report examines the performance of various classification algorithms (IBk, J48, AdaBoostM1, and NaiveBayes) and clustering methods (SimpleKMeans) on different datasets using WEKA. We evaluate classification accuracy and clustering Sum of Squared Errors (SSE) to assess the effectiveness of each method.

1 Introduction

The purpose of this lab is to compare the performance of several common classification and clustering methods using WEKA. We evaluate IBk (k-Nearest Neighbors), J48 (Decision Tree), AdaBoostM1 with J48, and NaiveBayes for classification tasks and SimpleKMeans for clustering tasks. Results are analyzed in terms of prediction accuracy for classification and SSE for clustering.

2 Part I: Classification Task

2.1 Methodology

We use the following datasets: *letter.arff*, *segment.arff*, and *waveform-5000.arff*. Each dataset is processed using IBk, J48, AdaBoostM1, and NaiveBayes with 10-fold cross-validation. The key parameters for each classifier are as follows:

- **IBk**: K values set to 1, 3, and 5.
- **J48**: Minimum number of instances per leaf (M) set to 2 and 4.
- **AdaBoostM1 (J48)**: Base classifier J48 with M=2.
- **NaiveBayes**: All parameters set to default.

2.2 Results

2.2.1 IBk Classification

Table 1 presents the accuracy of IBk classification for K values 1, 3, and 5 across the three datasets.

Table 1: Accuracy of IBk Classification

Dataset	**K=1**	**K=3**	**K=5**
letter	96.03	95.62	95.52
segment	97.14	96.02	95.06
waveform-5000	73.62	77.7	78.94

2.2.2 J48 and AdaBoostM1 Classification

Table 2 summarizes the accuracy results for J48 with M values of 2 and 4, and AdaBoostM1 with J48 (M=2) as the base classifier.

Table 2: Accuracy of J48 and AdaBoostM1 Classification

Dataset	**J48 (M=2)**	**J48 (M=4)**	**AdaBoostM1+J48 (M=2)**
letter	87.98	86.56	95.54
segment	96.93	96.06	98.53
waveform-5000	75.08	75.82	80.68

2.2.3 NaiveBayes Classification

Table 3 provides the accuracy results for NaiveBayes classification on each dataset.

Table 3: Accuracy of NaiveBayes Classification

Dataset	**Accuracy (%)**
letter	64.12
segment	80.22
waveform-5000	80.00

3 Part II: Clustering Task

3.1 Methodology

For the clustering task, we use SimpleKMeans to cluster each dataset with specified K values. We use the training set and evaluate each K-means model by measuring the Sum of Squared Errors (SSE). The chosen K values for each dataset are:

- **letter**: K = 11, 24, 38
- **segment**: K = 3, 5, 10
- **waveform-5000**: K = 2, 3, 5

3.2 Results

Table 4 shows the SSE results for SimpleKMeans clustering for each dataset with different K values.

Table 4: SSE of SimpleKMeans Clustering

Dataset	**K=K1**	**K=K2**	**K=K3**
letter	16513.56	9824.97	5132.69
segment	2173.14	1296.73	759.97
waveform-5000	5465.74	3846.65	3432.81

4 Conclusion

This report evaluates classification and clustering methods on three datasets using WEKA. IBk, J48, AdaBoostM1, and NaiveBayes are tested for classification, with AdaBoostM1 using J48 decision trees being the most accurate model across all datasets. SimpleKMeans clustering SSE is analyzed for different K values where the lowest error values are associated with the highest k values for each dataset.