CS584 – Data Mining

CA4 – Experimenting with Weka

Name: Prashanna Raj Pandit & Asha Shah\_\_\_\_\_ Grade:\_\_\_\_\_\_

This assignment is designed to get you started with Weka machine learning workbench (<https://www.cs.waikato.ac.nz/ml/weka/index.html>). Specifically, you will experiment with some simple algorithms such as OneR, Naïve Bayes, PRISM, and ID3. For a comprehensive description of Weka, refer to relevant parts in Witten 4. The following is a list of the tasks with activities and questions.

1. Read Witten appendix B and review chapter 4.
2. Download the most recent stable version of Weka 3.8 (stable version) and install it on your computer. (<https://waikato.github.io/weka-wiki/downloading_weka/>)
3. InstallsimpleEducationalLearningSchemes**package accessible from tools->package manager and include classes for ID3, NaiveBays, and PRISM.**  Note that if you do not have a computer, please email me or contact Greg Bartholomew ([gbartho@siue.edu](mailto:gbartho@siue.edu) ) immediately to have the software installed on our lab computers.
4. Use Weka explorer to determine the best One Rule (OneR) for weather (weather.nominal.arff) and soybean data sets. Include screenshots of the best one rule generated from each experiment and respective error rates. For this experiment, use the entire data set for both training and testing (Test Option 1). (30 points)

**🡺**

**OneR for Weather.nominal.arff**

**A screenshot of a computer

Description automatically generated**

**OneR for Soybean Dataset**

**A screenshot of a computer

Description automatically generatedA screenshot of a computer code

Description automatically generated**

1. OneR Algorithm Results

**Weather.nominal.arff Dataset**

* **Best Rule**: outlook = sunny → no; outlook = overcast/rainy → yes
* **Accuracy**: 71.43% (10/14 correct)
* **Error Analysis**:
  + Misclassified 4 instances (2 false positives, 2 false negatives).
  + The rule is simple but struggles with rainy/sunny edge cases.

**Soybean Dataset**

* **Best Rule**: Uses fruit-spots attribute (e.g., absent → alternarialeaf-spot).
* **Accuracy**: 40.85% (279/683 correct).
* **Error Analysis**:
  + Very low accuracy due to soybean's complexity (19 classes, 36 attributes).
  + OneR’s single-rule approach is too simplistic for this dataset.

**Key Insight**:  
OneR works well for simple datasets (like weather) but fails for complex, multi-class problems like soybean, where a single rule cannot capture nuanced patterns.

1. Use Weka explorer to apply Naïve Bayes classification to the same soybean dataset used in previous part. For this test, we are going to run the classification algorithm two times. We will use training set for testing the first time and 10-fold Cross Validation (CV) the second time while keeping all other setting as default. Include screenshot to display results for the two runs. (30 points)

🡺

Naïve Bayes for Soybean Dataset

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

10-fold Cross Validation (CV)

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

**Training Set Evaluation**

* **Accuracy**: 93.70% (640/683 correct).
* **Strengths**:
  + High precision/recall for most classes (e.g., 100% for diaporthe-stem-canker).
  + Handles multi-class problems better than OneR.
* **Weaknesses**:
  + Overfitting risk (tested on training data).
  + Lower performance for frog-eye-leaf-spot (72.5% recall).

**10-Fold Cross-Validation**

* **Accuracy**: 92.97% (635/683 correct).
* **Comparison**:
  + Slightly lower accuracy than training evaluation, confirming minor overfitting.
  + Robust performance across folds (generalizes well).

**Key Insight**:  
Naïve Bayes is highly effective for soybean data, with consistent cross-validation results. The small accuracy drop (~0.7%) suggests minimal overfitting.

1. Use Weka explorer to use Prism to generate rules for the weather.nominal.arff data. Run Prism with training set and include screenshots displaying results. (25 points)

🡺

A screenshot of a computer

Description automatically generated

**Rules Generate**

If outlook = overcast → yes

If humidity = normal and windy = FALSE → yes

If temperature = mild and humidity = normal → yes

If outlook = rainy and windy = FALSE → yes

If outlook = sunny and humidity = high → no

If outlook = rainy and windy = TRUE → no

* **Accuracy**: 100% (perfect classification).
* **Analysis**:
  + PRISM created a rule set covering all edge cases (e.g., rainy+windy → no).
  + More nuanced than OneR, leveraging multiple attributes.

**Key Insight**:  
PRISM’s rule-based approach outperforms OneR for the weather dataset, achieving perfect accuracy by combining conditions.

**Comparative Discussion**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Algorithm** | **Dataset** | **Accuracy** | **Strengths** | **Limitations** | | --- | --- | --- | --- | --- | | **OneR** | Weather | 71.43% | Simple, interpretable | Low accuracy for complex rules | |  | Soybean | 40.85% | – | Fails with multi-class data | | **Naïve Bayes** | Soybean (Train) | 93.70% | Handles multi-class, high recall | Risk of overfitting (minor) | |  | Soybean (10-CV) | 92.97% | Generalizes well | Slightly lower than training | | **PRISM** | Weather | 100% | Perfect accuracy, clear rules | Rules may overfit small datasets | |

**Conclusions**

1. **OneR** is suitable for binary/monotonic problems (e.g., weather) but inadequate for complex datasets.
2. **Naïve Bayes** excels in multi-class scenarios (soybean) with robust cross-validation performance.
3. **PRISM** generates interpretable rules and achieves perfect accuracy but may overfit on tiny datasets like weather (14 instances).

**Recommendation**:  
For soybean data, Naïve Bayes is the best choice due to its balance of accuracy and generalizability. For simpler datasets, PRISM or OneR may suffice if interpretability is prioritized.