# Feature Subset Selection for Rule Induction Using RIPPER


**Jihoon Yang**
Computer Science Dept.
Iowa State University
Ames, IA 50011
yang@cs.iastate.edu

**Asok Tiyyagura**
Computer Science Dept.
Iowa State University
Ames, IA 50011
asokt@cs.iastate.edu

**Fajun Chen**
Computer Science Dept.
Iowa State University
Ames, IA 50011
fjchen@cs.iastate.edu

**Vasant Honavar**
Computer Science Dept.
Iowa State University
Ames, IA 50011
honavar@cs.iastate.edu



## Abstract

Many existing rule learning systems perform poorly on large noisy datasets because of the presence of irrelevant features. In this paper we propose a hybrid approach combining Genetic Algorithms and RIPPER, to design better classifiers. Our experiments with several benchmark datasets show the feasibility of this approach.


The choice of features used to represent patterns has a strong impact on the accuracy of the classifier, the number of examples needed to attain a given classification accuracy on test data, the cost of classification, and the comprehensibility of the learned classifier. Feature subset selection [1] involves selecting a subset of features from a much larger candidate set of features so as to optimize multiple criteria such as the accuracy and the cost of pattern classification. RIPPER[2] is a flexible rule learning algorithm which has been demonstrated to perform well on a number of datasets. It has a time complexity of $O(m log^2 m)$ (where m is number of samples). RIPPER's relatively high accuracy, speed, and the simplicity and comprehensibility of the resulting rules make it attractive for pattern classification applications. In GA-RIPPER, each candidate feature subset is represented by a binary vector of dimension $m$ (the number of features). If a bit is a 1 (0), it means that the corresponding feature is selected (dropped). The test accuracy of rules learned by RIPPER using a given feature subset is used to measure the fitness of the feature subset. The speed of RIPPER makes its use feasible for use in feature selection using GA where a large number of fitness evaluations are needed.

The goal of our experiments was to compare the classification accuracy of RIPPER with and without feature subset selection on several datasets from UCI-Machine Learning Repository as well as some benchmark data for the document classification task. The GA-RIPPER implementation used the Tournament selection strategy with population size of 50. Each run consisted of 20 generations with probability of crossover and probability of mutation set to 0.6 and 0.001 respectively (based on results of several preliminary runs).

Table 1: The table compares the mean and std. deviation of the accuracies obtained using all the features (by RIPPER) with those obtained using feature selection (by GA-RIPPER), using 10-fold crossvalidation among the 5 independent runs of the genetic algorithm. The total number of available features and the number of selected features are denoted by Tot. and Sel. respectively.

| | RIPPER | | GA-RIPPER (average) | |
|---|---|---|---|---|
| *Dataset* | *Tot.* | *Accuracy* | *Sel.* | *Accuracy* |
| **Ionosphere** | 34 | 88.91 ± 1.4 | 14.6 ± 4.1 | 94.1 ± 0.7 |
| **Promoters** | 57 | 84.75 ± 5.0 | 28 ± 3.5 | 92.8 ±1.4 |
| **Sonar** | 60 | 76.86 ± 2.4 | 30.4 ± 2.7 | 81.24 ±2.3 |
| **Vehicle** | 18 | 67.25 ± 1.7 | 9.4 ± 1.3 | 73.24 ± 0.1 |
| **Votes** | 16 | 94.0 ± 1.0 | 4.8 ± 1.3 | 96.1 ±0.3 |
| **Wine** | 13 | 91.95 ± 1.9 | 7.4 ± 0.5 | 96.95 ±0.5 |
| **Zoo** | 16 | 89.00 ± 4.2 | 8.5 ± 1.9 | 95.2 ±0.8 |
| **Abstract1** | 790 | 84.0±2.8 | 385.8 ± 17.4 | 87.0 ± 0.0 |
| **Abstract2** | 790 | 86.00±2.8 | 402.4 ± 6.8 | 88.0 ± 0.0 |
| **Reuters1** | 1568 | 92.77±0.9 | 787 ± 23.5 | 97.4 ± 0.9 |
| **Reuters2** | 435 | 81.80±1.8 | 214.6 ± 7.4 | 92.24 ± 0.5 |
| **Reuters3** | 1440 | 95.92±0.7 | 719.2 ± 19.4 | 97.99 ± 0.1 |

Results indicate that the GA-selected subset of features resulted in substantial improvement in the classification accuracy. Furthermore, an examination of the learned rules showed that the rules were generally simple and easy to comprehend. We conclude that genetic algorithms offer an attractive approach to feature subset selection in data-driven rule induction. Several practical applications (e.g., diagnosis, automated data mining and knowledge discovery, power system security assessment, sensor system design in robotics and control, intrusion detection in distributed systems) can benefit from this approach to the design of simple, efficient, robust, and comprehensible pattern classifiers. Multi-objective fitness functions that take into account several design criteria (e.g., accuracy, cost of classification, etc.) can be easily incorporated into GA-RIPPER. Work in progress is aimed at extensive experimental evaluation of GA-RIPPER and its variants on a number of real-world problems.

[1] Motoda & Liu (1998) Feature Extraction, Selection and Construction: A Data Mining Perspective, Kluwer.

[2] W. Cohen (1995). In: Proc. ICML, Morgan Kaufmann