Discovering Rules in the Poker Hand Dataset

Robert Cattral
Carleton University
1125 Colonel By Drive
Ottawa, On, Canada K1S 5B6
cattral@gmail.com

Franz Oppacher
Carleton University
1125 Colonel By Drive
Ottawa, On, Canada K1S 5B6
oppacher@scs.carleton.ca

Categories and Subject Descriptors
I.2.6 [**Artificial Intelligence**]: Learning – *concept learning, knowledge acquisition.*

General Terms
Algorithms

Keywords
Data mining, classification, dataset, automatic rule discovery, genetic algorithms, hybrid algorithm.

1. INTRODUCTION
In recent years the popularity of Data Mining systems has increased substantially, and with that an interest in developing new algorithms for knowledge discovery. Testing machine learning algorithms is a non-trivial task that requires empirical studies be conducted, and the results compared against other published work.

The performance of a knowledge discovery technique is often evaluated by mining several datasets and examining the predictive model that is generated for each one. The characteristics of a dataset will often make obvious the strengths or weaknesses of a learning algorithm, and a collection of substantially different datasets is a good gauge of overall performance. Data archives, such as the UCI Machine Learning Repository [1], exist to facilitate this type of testing.

2. THE POKER HAND DATASET
The *Poker Hand* dataset was generated with the intention that it be difficult to discover the underlying concepts yet easy to analyze them. Because of the way the problem is represented it is difficult to discover rules that can correctly classify poker hands, although the simple nature of the game makes it trivial for the human analyst to validate potential rules objectively. The solution space is bounded because there are a finite number of valid poker hands. A valid hand is restricted to five unique cards in any position drawn from a standard deck of 52 playing cards.

Each hand consists of five cards that are not sorted by suit or rank, which means that the entire dataset contains approximately 311.8 million instances.

GECCO'07, July 7–11, 2007, London, England, United Kingdom
ACM 978-1-59593-697-4/07/0007.

Each instance consists of five cards that are not sorted by suit or rank, followed by the class. There are a total of ten predictive attributes that describe the five cards, as follows: {Suit of card #1, Rank of card #1, …, Suit of card #5, Rank of card #5}.

3. EXPERIMENTS USING WEKA
For the purpose of determining how difficult this classification problem is, we conducted an empirical study of several algorithms using the *Waikato environment for knowledge analysis* (Weka [2]).

The goal of a learning algorithm is to achieve better than 50.121%, which is the percentage of instances that fall within the majority (*nothing in hand)*) class. An algorithm that has predictive accuracy lower than this is worse than always choosing the majority, while values higher than this indicate that some correct predictions are being made.

Experiments with algorithms in Weka showed that most achieved 50.121%. There were three algorithms that score better than this, with the best being JRip having a predictive accuracy of 56.616%.

4. EXPERIMENTS USING RAGA
Using a population size of 100 rules, RAGA [3] was able to achieve an average of 99.56% accuracy over 10 runs, with the best run scoring 99.76% correct. The best case in 10 runs rose to 99.96% when using a population size of 1000 rules.

During experiments RAGA discovered several 100% confident, full coverage rules for the poker hand dataset. Each of these rules answers an entire class with 100% confidence, and does not misclassify any instances belonging to other classes. The classes for which these rules exist are: *One pair*, *two pairs*, *three of a kind*, *full house*, *four of a kind*, and *straight flush*. Utilizing all of these rules would ensure that at least 49.29% of the entire domain is properly classified without any risk of misclassification.

5. REFERENCES
[1] C.L. Blake and P.M. Murphy. The UCI machine learning repository. In Irvine, CA: University of California, Department of Information and Computer Science, 1998. <http://www.cs.uci.edu/mlearn/MLRepository.html>

[2] Ian H. Witten, Eibe Frank. Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[3] Cattral, R. RAGA: Rule Acquisition With A Genetic Algorithm. Masters thesis, Carleton University, Ottawa, Ontario, Canada, 2001. ISBN: 0612577589