An Introduction to the WEKA Data Mining System

Zdravko Markov
Central Connecticut State University
New Britain, CT, USA
01-860-832-2712
markovZ@ccsu.edu

Ingrid Russell
University of Hartford
West Hartford, CT, USA
01-860-768-4191
irussell@hartford.edu

ABSTRACT

This is a proposal for a half day tutorial on Weka, an open source Data Mining software package written in Java and available from www.cs.waikato.ac.nz/~ml/weka/index.html. The goal of the tutorial is to introduce faculty to the package and to the pedagogical possibilities for its use in the undergraduate computer science and engineering curricula. The Weka system provides a rich set of powerful Machine Learning algorithms for Data Mining tasks, some not found in commercial data mining systems. These include basic statistics and visualization tools, as well as tools for pre-processing, classification, and clustering, all available through an easy to use graphical user interface.

Data Mining studies algorithms and computational paradigms that allow computers to discover structure in databases, perform prediction and forecasting, and generally improve their performance through interaction with data. Machine learning is concerned with building computer systems that have the ability to improve their performance in a given domain through experience. Machine learning and Data Mining are becoming increasingly important areas of engineering and computer science and have been successfully applied to a wide range of problems in science and engineering. Recently, acknowledging the importance of these areas in computer science and engineering, more work is being done to incorporate these areas into the undergraduate curriculum.

Weka is a widely used package that is particularly popular for educational purposes. It is the companion software package of the book "Data Mining: Practical Machine Learning Tools and Techniques" by Ian H. Witten and Eibe Frank. The Weka team has been recently awarded with the 2005 ACM SIGKDD Service Award for their development of the Weka system, including the accompanying book. As Gregory Piatetsky-Shapiro writes in the news item about this event (KDnuggets news, June 28, 2005), "Weka is a landmark system in the history of the data mining and machine learning research communities, because it is the only toolkit that has gained such widespread adoption and survived for an extended period of time (the first version of Weka was released 11 years ago)".

The purpose of this tutorial is to present an introduction to the Weka system and outline the major approaches to using Weka for teaching Machine Learning, Data and Web Mining. We will also present our experiences using Weka as a main tool for implementing Machine Learning and Web Mining student projects that have been developed in the framework of a National Science Foundation grant. In this framework, two basic applications of Weka will be used to illustrate the various Weka topics presented:

- Web document classification. Some basic classification schemes provided by Weka (Nearest Neighbor, Naïve Bayes and Decision trees) are used to create models of web documents in topic directories and then to classify new documents according to their topic.
- Intelligent web browser. Web documents are labeled with the preferences of web users and ML models are created. These models are then used to classify documents returned by web searches according to the user preferences.

In this framework the following topics will be covered:

- Data preprocessing and visualization
- Attribute selection
- Association rules
- Classification algorithms (OneR, Decision trees, Covering rules)
- Prediction algorithms (Naïve Bayes, Nearest neighbor, Linear models)
- Evaluation techniques
- Clustering (K-means, EM, Cobweb)

For each of these topics, examples of using Weka will be presented. No background in machine learning or data mining is needed.

Categories and Subject Descriptors

K.3.2 [**Computers and Education**]: Computer Science Education

General Terms: Experimentation

Keywords: Artificial Intelligence, Projects

REFERENCES

[1] Kumar, A., Kumar, D., Russell, I., "Non-Traditional Projects in the Undergraduate AI Course", *Proceedings of the Thirty-Seventh SIGCSE Technical Symposium on Computer Science Education,* ACM Press, New York, NY, February 2006.

[2] Markov, Z., Russell, I., Neller, T. *Proceedings of the Thirty-Fifth Annual Frontiers in Education Conference*, IEEE Press, October 2005.

[3] Mitchell, T., Does Machine Learning Really Work, *AI Magazine*, Vol. 18, No. 3, AAAI Press, Fall 1997.

[4] Neller, T., Presser, C., Russell, I., Markov, Z., "Pedagogical Possibilities for the Dice Game Pig", *The Journal of Computing Sciences in Colleges*, 21(5), May 2006.

ITiCSE'06, June 26–28, 2006, Bologna, Italy.
ACM 1-59593-055-8/06/0006.

[5] Neller, T., Markov, Z., Russell, I., “Clue Deduction: Professor Plum Teaches Logic”, *Proceedings of the International FLAIRS Conference*, AAAI Press, May 2006.

[6] Russell, I., Markov, Z., Neller, T., “Unifying an Introduction to Artificial Intelligence Course through Machine Learning Laboratory Experiences”, *Proceedings of the 2005 Annual American Society for Engineering Education Conference*, June 2005.

[7] Russell, S., J. and Norvig, P., *Artificial Intelligence: A Modern Approach*, Upper Saddle River, NJ: Prentice-Hall, second edition, 2002.

[8] Witten, I.H. and frank, E., *Data Mining: Practical Machine Learning Tools and techniques with Java Implementations*, Morgan Kaufmann, 1999.