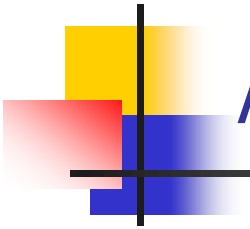
In Collaboration With

Machine Learning as a Contextualized Approach for Improving Student Engagement and Learning

Ingrid Russell (irussell@hartford.edu)
University of Hartford, CT, USA
Zdravko Markov, Central Connecticut State University, CT, USA
Susan Imberman, College of Staten Island, NY, USA



Abstract

Machine learning is providing methodology and technology for enhancing real-world applications in a wide range of areas and in several commercial applications. We present a multi-institutional project for teaching core Artificial Intelligence (AI) topics through a unifying theme of machine learning. A total of 26 adaptable, hands-on laboratory projects have been developed that can be closely integrated into introductory AI courses. This is a collaborative effort involving several faculty members working on the development of the material and also on its implementation and testing. Each project involves the development of a machine learning system in a specific application. The applications span a large area including network security, recommender systems, game playing, web document classification, vision, bioinformatics, pattern recognition, and data mining.

Project MLeXAI Goal

- The goal is to develop an adaptable framework for the presentation of core AI topics through a unifying theme of machine learning. A suite of hands-on term-long projects are developed, each involving the design and implementation of a learning system that enhances a commonly-deployed application.
- Applications span a large area including network security, recommender systems, game playing, robotics, cryptography, computer vision, web document classification, data integration in databases, bioinformatics, pattern recognition, and data mining.

Project MLeXAI Objectives

- Enhance student learning experience in AI by implementing a unifying theme of machine learning to tie together core AI topics.
- Increase student interest and motivation to learn AI by providing a framework for the presentation of the major AI topics that emphasizes the strong connection between AI and computer science.
- Highlight the bridge that machine learning provides between AI technology and modern software engineering.
- Introduce students to an increasingly important research area, thus motivating them to pursue further study in this area.

Features of MLeXAI Projects

- Teaching AI with hands-on experiments.
- Common features in different AI fields are unified through the theme of machine learning.
- Emphasis on application of ideas through implementation with implementation of concepts being central to the learning process.
- Design and implementation of learning systems.
- Practical approach that includes real-world applications.
- Easily adaptable and customizable.
- Various emphases, backgrounds and prerequisites that can serve different goals within the general framework of teaching AI.

Sample MLeXAI Projects

- Web Document Classification: Investigates the process of tagging web pages using a topic directory structure and applies machine learning techniques for automatic tagging.
- Data Mining for Web User Profiling Using Decision Tree Learning: Focuses on the use of decision tree learning to create models of web users.
- Character Recognition Using Neural Networks: Involves the development of a character recognition system based on a neural network model.

Sample Project: Web Document Classification

Goal

To investigate the process of tagging web pages using the topic directory structure and apply machine learning techniques for automatic tagging. www.dmoz.org (The Open Directory Project)

Sample Project: Web Document Classification

Project Phases

- Data collection – collecting a set of 100 web documents grouped by topic. Will serve as training set.
- Feature extraction and data preparation – web documents will be represented by feature vectors, which in turn are used to form a training dataset for the Machine Learning stage.
- Machine learning – applying learning algorithms to create models of the datasets. Using these models the accuracy of the initial topic structure is evaluated and new web documents are classified into existing topics.
- Analysis – identifying relations between approaches used in the project and AI areas of search and knowledge representation and reasoning.

Sample Project: Web Document Classification

Phase I – Data Collection

- Use a topic directory structure such as dmoz.org.
- Identify topics, each of which is well represented by a set of documents. Identify 20 web documents for each of the 5 topics.
- As these documents will be used for learning and classification experiments at a later stage, they must form a specific structure (part of the topic hierarchy).
- Each page must have enough keywords that represent the topic.
- Good to have topics at different levels of the topic hierarchy and with different distances from them.
- Could also use a Web crawler to collect the Web pages. WebSPHINX (free open source), a customizable Web crawler available at
  http://www-2.cs.cmu.edu/~rcm/websphinx/

Sample Project: Web Document Classification

Phase I – Data Collection: Topic Structure

Topic 1 *Topic 2* *Topic 3*
Computers: Artificial Intelligence: ***Machine Learning***
Topic 1 *Topic 2* *Topic 4*
Computers: Artificial Intelligence: ***Agents***
Topic 1 *Topic 5* *Topic 6*
Computers: *Algorithms*: ***Sorting and Searching***
Topic 1 *Topic 7* *Topic 8*
Computers: Multimedia: ***MPEG***
Topic 1 *Topic 9*
Computers: ***History***

Sample Project: Web Document Classification

Phase II – Feature Extraction and Data Preparation

- Web documents will be represented by feature vectors, to be used to form a training dataset for the machine learning stage.
  - Keyword (term) selection based on term frequencies.
  - Creating feature vectors – Boolean-based.
  - Creating data files for the Weka ML system.

Sample Project: Web Document Classification

1		2		3	
Words	Frequency	Words	Frequency	Words	Frequency
and	10	the	97	the	36
you	6	of	45	of	31
to	5	and	34	to	17
is	5	is	31	in	16
in	5	to	27	n	13
code	4	1	26	Flashsort	12
Visual	4	a	25	is	12
are	4	data	23	a	9
for	4	63	21	k	9
section	4	in	21	A	9
of	4	27	20	The	8
Basic	4	64	20	L	7
The	4	9	20	time	7
a	4	number	19	be	6
algorithms	4	14	19	permutatior	6
C	3	72	19	and	6
that	3	58	19	elements	6
sorting	2	sort	18	cycle	6
dictionaries	2	be	18	m	6
this	2	pass	15	vector	5
Source	2	The	15	for	5
may	2	it	15	by	5
be	2	It	15	algorithm	5
This	2	with	14	an	5
language	2	performanc	12	in-situ	5
source	2	sorting	12	are	4
structures	2	algorithm	12	new	4
files	2	then	11	algorithm	4
with	2	sorted	11	that	4
document	2	that	11	Neubert	4
programmin	2	are	11	page	3

Sample Project: Web Document Classification

Phase II – Feature Extraction and Data Preparation

- Creating the vector space model, a 100x100 matrix which serves as a representation of the 100 documents.
- Creating a dataset in the ARFF format (Attribute-Relation File Format) to be loaded into Weka.
- Weka Data Mining System is open source machine learning software package written in Java and available at: www.cs.waikato.ac.nz/~ml/weka/

Sample Project: Web Document Classification

Project Phase II – Feature Extraction and Data Preparation

```
,0,1,0,1,0,0,1,1,0,1,0,0,0,0,1,0,1,0,0,CompHistory
,0,0,0,0,0,0,1,1,0,0,1,0,0,0,1,0,1,1,0,CompHistory
,1,1,0,1,1,1,1,1,0,0,1,0,1,1,1,0,1,1,1,CompHistory
,0,1,0,1,0,0,1,1,0,0,1,0,0,0,1,0,1,0,1,CompHistory
,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,1,0,Search_Sort
,1,1,0,0,0,1,1,0,0,0,1,0,1,0,0,0,0,0,0,Search_Sort
,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,0,Search_Sort
,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,Search_Sort
,0,1,0,0,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,Search_Sort
,0,1,0,0,0,1,0,0,0,0,1,0,1,0,0,0,0,0,0,Search_Sort
,0,1,0,0,0,0,0,0,0,0,1,0,1,0,1,0,0,0,0,Search_Sort
```

Sample Project: Web Document Classification

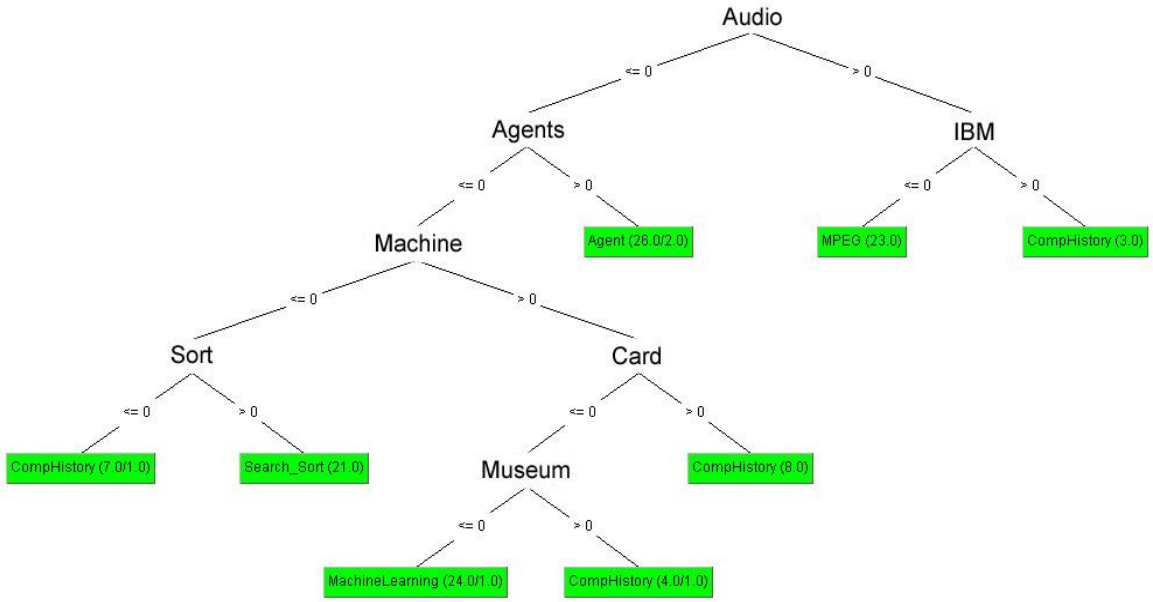
Phase III – Machine Learning

- Preprocessing and analyzing the data created in Phase II.
- Creating models of the datasets.
- Using Weka's decision tree algorithm (J48) to examine the decision tree generated from the dataset.
- Analyzing the decision trees generated with different datasets.
- New web document classifications.

Sample Project: Web Document Classification

Phase III – Machine Learning

WEKA used the supplied ARFF file to develop the most accurate decision tree for predicting the topics of each web page as represented in the feature vector.



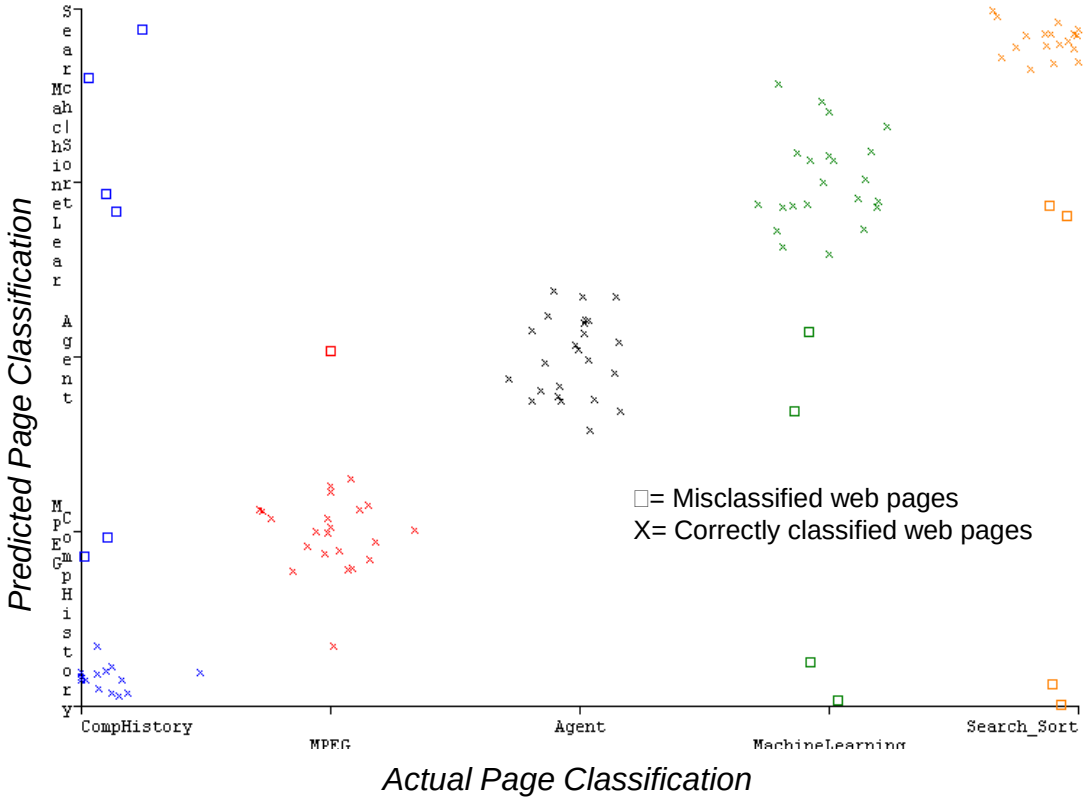
The J48 Decision Tree correctly classified 87.069% of the pages using a 10-Fold Cross Validation.

Results

===== Stratified Cross-Validation =====

Correctly Classified Inst.(101)	87.069 %
Incorrectly Classified Inst.(15)	12.931 %
Kappa statistic	0.8379
Mean absolute error	0.0633
Root mean squared error	0.2245
Relative absolute error	19.7946 %
Root relative squared error	56.1278 %
Total Number of Instances	116

Error Analysis Plot



Conclusion

- Results of assessment were very positive and showed that students had good experiences in the classes.
- While covering the main AI topics, the course provided students with an introduction to and an appreciation of an increasingly important area in AI, Machine Learning.
- Using a unified theme was helpful and motivating for the students. Students saw how simple search programs evolve into more interesting ones, and finally into a learning framework with interesting theoretical and practical properties.

Acknowledgement

This work is supported in part by National Science Foundation grant DUE-0409497 and DUE-0716338.

Additional Information

- MLeXAI website: http://uhaweb.hartford.edu/compsci/ccli
- Contact: Ingrid Russell irussell@hartford.edu