Zdravko Markov, Ph.D.
Professor of Computer Science
Central Connecticut State University

Phone: (860) 832-2711
E-mail: markovz@ccsu.edu
URL: http://www.cs.ccsu.edu/~markov/

Developing a Data Mining Software Suit for Attribute Selection and Clustering

Sabbatical Project Fall-2013

Abstract

The purpose of this project is to enhance teaching Machine Learning and Data/Web Mining courses at CCSU and elsewhere, and to contribute to the academic field in general by creating a Data Mining software suite for attribute selection and clustering. The software suite will extend the companion web site of the book "*Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*" (While, 2007) written during the previous sabbatical project, and thus will provide free access for faculty and students to algorithms, technical documentation, examples and laboratory exercises to support courses and research in these areas. The results from the analysis and the comparative study of the algorithms included in the software suite will be published in a major conference or journal and thus will make a contribution to the research in the field.

Statement of purpose and objective

In Machine Learning and Data Mining data are usually represented as a set of features (attributes). The choice of this set is critical for the quality and efficiency of the overall learning/mining process. Therefore attribute selection is a fundamental part of the research and algorithm development in these areas. Clustering is a technique for grouping objects by similarity and finding patterns in data, which heavily depends on the proper choice of attributes. Therefore algorithms for attribute selection and clustering are an intrinsic part of any Machine Learning or Data Mining System and also major topics in any course or book covering these areas.

The purpose of this project is to enhance teaching Machine Learning and Data/Web Mining courses at CCSU and elsewhere, and to contribute to the academic field in general by creating a Data Mining software suite for attribute selection and clustering. It will include implementations of popular existing algorithms and two original algorithms that I developed during my research reassigned time project in Fall-2011. The software suite will provide user-friendly access for faculty and students to the algorithms, technical documentation, examples and laboratory exercises to support courses and research in these areas. All these resources will be used to

extend the companion web site of the book "*Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*" (While, 2007) written during my previous sabbatical project. This extension will complement the material of the book thus supporting courses based on that book, and also other courses in the area of Machine Learning and Data/Web Mining. The laboratory projects will be used to extend the collection of Machine learning teaching resources developed in the framework the project "Machine Learning Experiences in Artificial Intelligence: A Multi-Institutional Project", funded by an NSF grant during the period 2007-2009, for which I was a co-PI. The results from the analysis and the comparative study of the algorithms included in the software suite will be published in a major conference or journal and thus will make a contribution to the research in the field.

Description of existing knowledge and/or work to date related to the project

Machine Learning, Data Mining and Web Mining are rapidly growing IT areas with large impact on society – many businesses rely on Data Mining for data modeling and prediction, and the search engines using Web Mining techniques provide wide access to the world's largest knowledge repository, the Web. Machine Learning provides most of the theoretical foundation for both. All this has led to increasing demand for Computer Scientists trained in these areas. Following this trend the CS department now is offering three graduate courses (part of the CIT and Data Mining programs in CCSU) covering Machine Learning, Data Mining and Web Mining and is also introducing more Machine Learning related content in some undergraduate courses. I have been developing and teaching three graduate courses in these areas: CS570 - Topics in AI: Machine Learning, CS580 - Topics in Data Bases: Data Mining, and CS580 - Topics in Data Bases: Web Mining. I've also taught CS 463 – Artificial Intelligence, where I included a large section of Machine Learning. When I developed and taught these courses I tried to meet the educational objectives of the CS department as well as the CCSU's comprehensive educational strategy. Therefore I put the emphasis on the Computer Sciences and Mathematical foundations and introduced more independent studies and research oriented projects. All course materials are available directly on the web or through the Blackboard Vista course management system:

- Web Mining at http://www.cs.ccsu.edu/~markov/ccsu_courses/WebMining.html. This web page includes class notes and detailed description of three projects covering the basic areas of Web Mining: Web Crawler, Web Document Classification and Intelligent Web Browser.
- Machine Learning course syllabus at http://www.cs.ccsu.edu/~markov/ccsu_courses/CS570Syllabus.html. The class notes are available in BB Vista and also include:
  - A link to comprehensive lecture notes on Machine Learning that I wrote for this course (66-page document in PDF format available at http://www.cs.ccsu.edu/~markov/ccsu_courses/lnml.pdf).
  - A set of Prolog programs that I implemented to illustrate the algorithms discussed in the course and provide tools to students to do their projects. These programs are indexed in a separate web page available at http://www.cs.ccsu.edu/~markov/ccsu_courses/mlprograms/.

- Data Mining course syllabus at http://www.cs.ccsu.edu/~markov/ccsu_courses/CS580Syllabus.html. The complete set of class materials is available in BB Vista and includes detailed lecture notes and 8 projects requiring independent study and practical work with Data Mining software.

During my previous sabbatical leave in 2005-6 I wrote a book on Web Mining [1] with my colleague from the Math department Dan Larose. The book was published by Wiley in 2007 and has been used since then for our courses. This project also resulted in data sets and lab exercises to accompany and enhance the book material. The suggested software to be used for this purpose was the Data Mining system Weka [2]. It provides a large collection of Machine Learning and Data Mining algorithms and is one of the most popular data mining systems for academic purposes. Two important topics covered in the book, but not well supported by Weka, are unsupervised attribute selection and clustering. During my work on the book I developed a new approach to attribute selection, which is described in Chapter 4, Section "MDL-Based Model and Feature Evaluation". Since then I've been working on implementations of this approach and applying it to clustering. The results of this work were summarized during my research reassigned time project in Fall-2011. Two algorithms were developed – MDLranker, which evaluates each data attribute and orders attributes accordingly, and MDLcluster that uses this evaluation for clustering data. These algorithms were implemented in Java and included in the companion web site of the book along with documentation and datasets [3]. More recently I did a comparative study and experimental evaluation of these algorithms, which showed that they outperform most of the well-known algorithms in this area. I wrote a paper [4] describing these results, which was submitted to the leading international machine learning conference ICML 2013 and is now under review. A copy of the paper is attached to this application.

Another related project, "Machine Learning Experiences in Artificial Intelligence: A Multi-Institutional Project" (http://uhaweb.hartford.edu/compsci/ccli/index.htm), was funded by a grant from NSF (DUE-0716338), for which I was a co-PI. One of the major results of this project was creating a repository with sample laboratory projects to enhance teaching Machine Learning and related topics in the Artificial Intelligence course. I developed three Machine Learning projects for this repository, all based on using Weka, the book [1] and materials from its companion web site: "Web User Profiling" (http://uhaweb.hartford.edu/compsci/ccli/samplep.htm), "Probabilistic Reasoning with Naïve Bayes and Bayesian Networks" (http://uhaweb.hartford.edu/compsci/ccli/pr.htm) and "Web Document Classification" (http://uhaweb.hartford.edu/compsci/ccli/wdc.htm).

Description of proposed sabbatical activities and/or methodology

The project will include three major steps. Taking into account my experience in evaluating and creating machine learning software I estimate that Step 1 and 3 can be accomplished in one month each, and Step 2 will take two months.

1. *Evaluating and choosing algorithms for attribute selection and clustering to be included in the software suite.* Two sources will be used for this purpose. The first one is the open source system Weka [2], one of the most popular data mining systems for academic

purposes. It's is written in Java, which makes it platform independent and easily adaptable for different teaching and research purposes. The second source is the two algorithms based on the attribute evaluation approach originally described in Chapter 4 of the book [1], Section "MDL-Based Model and Feature Evaluation". They were developed as a result of my research reassigned time project in Fall-2011. The first one (MDLranker) computes the MDL evaluation measure of each data attribute and orders attributes accordingly. The second algorithm (MDLcluster) uses the evaluation measure for clustering data. All algorithms chosen for inclusion in the software suit will be analyzed and compared by literature research and empirically evaluated on popular benchmark data sets.

2. *Creating a software suite implementing efficiently the selected algorithms and providing user-friendly access to them*. The system will be written in Java by using programs and parts of the Weka open source and the programs implementing the two original algorithms [4]. The existing implementations will be first optimized to improve their efficiency. Then they will be modified in order to use a common input data format and unified structure of the results. The Weka format for input data will be used for this purpose, which will allow using the large and widely available collection of datasets provided with the Weka system. The results will be presented also in the Weka format so that they may be easily analyzed and compared with results from other research and teaching projects based on the Weka system.
3. *Collecting/creating data sets, exercise and laboratory projects to support teaching courses and academic research in the areas of Machine Learning and Data/Web Mining by using the software suite*. Experiments with datasets from the large collection provided with the Weka system will be performed in order to find the datasets producing results that best illustrate the approaches to attribute selection and clustering. These datasets will be used as examples in the technical documentation and included in exercises and laboratory projects that will accompany the software suit.

Statement of potential value

The major expected result of this project is the development of a software suit, exercises and laboratory projects, which will become valuable resources for enhancing teaching Machine Learning, Data/Web Mining and related courses at CCSU. These resources will substantially extend the existing collection of resources accompanying the book "*Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*" (While, 2007). The book is used for teaching CS 580 - Web Mining, a course in the MS program in Computer Information Technology (CIT) and some courses in the MS Program in Data Mining. Independently from the book, the software, the exercises and the laboratory projects can be easily incorporated in two other graduate courses in the CIT program: CS 570-Machine Learning, and CS 580 – Data Mining, and can be also used to enhance the Machine Learning section of the CS 462 - Artificial Intelligence (taught both for the CS undergraduate program and for the MS CIT program).

As a part of the book companion web site the developed resources will be freely available for the wider academic community and thus will contribute to the teaching and research in the field by providing tools for analysis and empirical evaluation of algorithms for Machine Learning and Data Mining.

The results from the analysis and the experimental evaluation of the algorithms will be described in a scientific paper and submitted to a major Machine Learning or Data Mining conference (ICML, ECML, or ICDM), which will be an important contribution to my professional growth and to the field in general.

Statement of expected outcomes

This project will extend the results of my previous sabbatical project, the book "*Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*" (While, 2007) ", written in 2005/6. This book includes chapters on attribute selection and clustering. The project also resulted in data sets and lab exercises to accompany the book. The suggested software used for this purpose was the Data Mining system Weka. This system does not include any unsupervised attribute selection algorithms and provides a limited number of clustering algorithms. The software suite that will be developed for the current project will include such algorithms and extend the set of clustering algorithms. They will be made available online as part of the companion web site of the book along with data sets, exercises and laboratory projects to support courses based on that book and promote academic research in the field.

References

1. Zdravko Markov and Daniel T. Larose. *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*, Wiley, April 2007. Companion website at http://www.dataminingconsultant.com/DMW.htm. Lecture slides, data sets and software available at http://www.cs.ccsu.edu/~markov/.
2. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1, 2009. (http://www.cs.waikato.ac.nz/ml/weka/).
3. Zdravko Markov. Java Classes for MDL-Based Attribute Ranking and Clustering, Manual for the companion software for the book "*Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*, Wiley, April 2007", 2012 (http://www.cs.ccsu.edu/~markov/dmwsoftware.zip).
4. Zdravko Markov. MDL-Based Attribute Ranking and Clustering, Submitted to the 30th International Conference on Machine Learning (ICML 2013), Atlanta Georgia, June 16 - 21, 2013 (http://icml.cc/2013/).