Project EMD-MLR: Educational Material Development and Research in Machine Learning for Undergraduate Students

Georgios C. Anagnostopoulos[*], Michael Georgiopoulos[**], Ken Ports[*], Samuel Richie[**], Nelly Cardinale[***], Melinda White[****], Veton Kepuska[*], Phil K. Chan[*], Annie Wu[**], Marcella Kysilka[**]

(*) Florida Institute of Technology, Melbourne, FL / (**) University of Central Florida, Orlando, FL / (***) Brevard Community College, Melbourne, FL / (****) Seminole Community College, Orlando, FL

Abstract

This paper presents implementation and organizational details, as well as accomplished outcomes on a National Science Foundation funded project under the auspices of the Educational Materials Development track of the program entitled Course, Curriculum and Laboratories Improvement (CCLI-EMD). The title of the grant is "PROJECT EMD-MLR: Educational Materials Development through the Integration of Machine Learning Research into Senior Design Projects". The project partners are two major universities in Central Florida, Florida Institute of Technology (FIT) in Melbourne and the University of Central Florida (UCF) in Orlando. In addition to the two universities, there are two 2-year Central Florida colleges, Seminole Community College (SCC) in Oviedo and Brevard Community College (BCC) in Palm Bay.

Project EMD-MLR is a "proof-of-concept" project focused on Machine Learning (ML), whose immediate objectives are i) the development of educational material in the form of software implementations of ML algorithms and the compilation of accompanying documentation; the creation of an on-line ML Repository via the archiving of the developed products; the authoring of 3-4 chapters of an undergraduate ML textbook, ii) the introduction of ML research in the participating institutions' curricula and iii) the dissemination of the developed material through the on-line outlet, affiliated academic institutions and industries. The project's objectives are achieved by a team of faculty with strong teaching and research experience in ML, as well as extensive experience on supervising senior design teams. EMD-MLR's objectives will impact 80 students in a span of 2 years, some of which are members of underrepresented groups (in particular UCF and SCC students). Furthermore, the project's focus enhances partnerships amongst 4 neighboring institutions, and many more affiliate Universities dispersed within the US and abroad. The University partnerships constitute the anchor of an elaborate dissemination plan that is multi-faceted and self-sustained.

1. Introduction

This paper describes "PROJECT EMD-MLR: Educational Materials Development through the Integration of Machine Learning Research into Senior Design Projects", a prototype project pertaining to the Educational Materials Development track of NSF's Course, Curriculum and Laboratories Improvement program (CCLI-EMD). The project partners are two major universities in Central Florida, namely Florida Institute of Technology (FIT) in Melbourne and the University of Central Florida (UCF) in Orlando. In addition to the two universities, there are two 2-year Central Florida colleges, Seminole Community College (SCC) in Oviedo and Brevard Community College (BCC) in Palm Bay. The project is being funded under NSF grant #0341601 for the period of May 2003 to Arpil 2006.

The theme of EMD-MLR is Machine Learning (ML), an internationally growing and ever-advancing research discipline that uses principles of artificial intelligence, statistics and mathematics to address current and future technological challenges. It solves problems in many specialization areas such as pattern, speech, image recognition, data mining, robot vision and natural language understanding among others. Therefore, it is important to introduce ML starting at the undergraduate level (contrary to the traditional trends of reserving it for graduate students) in an appropriate form, which will invoke excitement among the undergraduate students, equip them with the basic background and prepare them to conduct research in ML at the graduate level.

Project EMD-MLR is a "proof-of-concept" project focused on Machine Learning (ML) and its immediate objectives are i) the development of educational material in the form of software implementations of ML algorithms and the compilation of accompanying documentation; the creation of an on-line ML Repository via the archiving of the developed products; the authoring of 3-4 chapters of an undergraduate ML textbook, ii) the introduction of ML research in the participating institutions' curricula and iii) the dissemination of the developed material through the on-line outlet, affiliated academic institutions and industries. Project EMD-MLR's objectives are presented analytically in Section 2, while Section 3 of the paper elaborates on the project's methodology. Section 4 underlines the merits of the project and, finally, in Section 5 we provide a brief summary.

2. Project Objectives

The project's primary goal is to introduce research into the undergraduate curriculum of many engineering and science disciplines by involving senior design teams in machine learning projects, an area that is of multi-disciplinary nature and is ever-developing. The idea of introducing research into the curriculum is not new, it is a national trend, and NSF has embraced this idea by supporting initiatives that foster this approach, such as the REU (Research Experiences for Undergraduates), CRCD (Combined Research and Curriculum Development), etc. As it was emphasized in Project Kaleidoscope (1999), which was funded by NSF, "the undergraduate years are the last opportunity for academic study of STEM (science, technology, engineering and math) subjects by many of the future leaders of our society—the executives, government officers, lawyers, clergy, journalists and others who will have to make momentous decisions involving science and technology". As a result, the effort of involving undergraduate

students in research could be viewed as a step in the right direction. The minimum number of undergraduate students that are affected by the project is 40 per year, 80 in total.

A second objective of EMD-MLR is to develop educational materials focused on Machine Learning, that will be of value to many academicians, students and professionals with interest in this research area, or applications that this field addresses. In particular, it is planned to develop: (i) a software collection of machine learning related algorithms based on MATLAB, a popular, scientific computational platform developed by Mathworks, with accompanying documentation and usage examples based on real-world problems. The project's ultimate intention is to create a Machine Learning repository at FIT that would be useful as an on-line resource dissemination tool to all machine learning researchers and students around the nation and all over the world. Although there is an abundance of machine learning code available in the public domain, these resources are mostly scattered among different sites and, therefore, are occasionally difficult to locate. Moreover, even well-known sites that serve as on-line Machine Learning repositories (such as the UCI-MLR at the University of California at Irvine) encompass heterogeneous, non-cohesive collections of code spanning different programming languages, software and user interface designs. In EMD-MLR the focus is in producing (or, sometimes, reproducing) well known and established machine learning algorithms and techniques within a common framework that will offer a unified user interface ("look-and-feel") to the code and its accompanying documentation. We are also in the process of producing a software development approach to facilitate the reuse and the embedding of the generated code by third parties, such as researchers and students. Over the prototype project's lifespan (2 years) an estimated 20 Machine Learning algorithms and approaches will be developed.

Additionally, we plan to develop educational material regarding selected, current topics in machine learning that will be used as chapters of an introductory textbook in Machine Learning for undergraduate students. Due to the mathematically and statistically rigorous nature of machine learning, most available textbooks in this field are geared towards the graduate level. However, the proposed textbook will built upon the proposed software/documentation collection and will be carefully designed in such a way that it will be appropriate for use in undergraduate EE, CpE and CS courses at both FIT and UCF. During the project's lifespan 3-4 chapters of the textbook will be developed.

Regarding the student recruiting, we involve students from 2-year community colleges in this educational development process. Community college students have limited opportunities to be involved in research. Through their involvement in the senior design or capstone classes these students will have the opportunity to appreciate research and its associated benefits and will be positively influenced to continue their education by pursuing a B.S. degree in a STEM (Science, Technology, Engineering or Mathematics) discipline. Senator Glenn has recently mentioned in one of his recent reports to Congress that there is a dire need to increase the number of STEM degrees. The proposed effort partially addresses this specific need. This kind of need was also stressed in separate reports (National Science Board, 1998 and Chronicle of Higher Education, 1999) where it is stated that a 44% increase in STEM jobs is projected in the 1996-2006 timeframe and three-fourths of this increase will happen in computer-related fields. The project involves 10 undergraduate students per year from the partnering community colleges, a total of 20 students over the project's duration. The recruiting process involves a careful screening of

student applications and the selection of the EMD-MLR participants relies on prior academic performance, and interpersonal skills, and focuses on creating a diverse participant body.

An integral component of the project is its dissemination plan. We will start the dissemination process of the educational materials by directing these dissemination efforts (as soon as the first year of this effort is completed) to a number of affiliated Universities, some of which serve a student population of underrepresented minorities. The institutions that are going to be directly benefiting from this EMD-MLR effort are the University of New Mexico (37% Hispanics in the EE program, 21% Hispanics in the CpE program, and 20% female student population in both programs), the United States Military Academy (a 4-year undergraduate institution with 4000 cadets of which 15% are female and 25% are minorities), and the University of Hartford (a 4-year undergraduate institution), among others. This addresses a national need to involve more students from underrepresented minorities in STEM activities. Based on national data (National Science Board, 1998, National Science Foundation, 1999) the percentage of Bachelor's degrees for underrepresented minorities (such as African-Americans, Hispanics, and women) remains well below their percentage in the population. These groups also account for less than 3.5% of Doctoral candidates, a number that has remained unchanged since 1976. This effort, through the active involvement of a targeted group of affiliate Universities, will address some of these needs.

3. Methodology

EMD-MLR's main approach is to involve undergraduate students into Student Design Project (SDP) teams with each team working on an individual machine learning project with a definite research orientation. The term "SDP" refers to senior design projects in engineering disciplines or capstone course in computer science disciplines. Each team consists of 4 undergraduate students (3 from a group of seniors at FIT or UCF and one sophomore student from BCC and SSC, respectively). Each SDP team is advised weekly by at least one EMD-MLR faculty and, if needed, will is more frequently by a Ph.D. student mentor. Each SDP consists of an (i) educational materials development component and a (ii) supervised research conduct component.

The former component entails the implementation and documentation of an assigned, Machine Learning algorithm and/or approach. Code is developed in the C++ programming language and executables are delivered as MATLAB MEX files. The design and implementation of the code follows well described, pre-specified guidelines, which are collectively referred to as development standards. These standards include software design, coding, and testing methodologies as well as applications of well established software engineering principles, which will guarantee the deployment of functionally robust, computationally and memory usage efficient code. Moreover, the development standards achieve a uniform "feel-and-look" across all team projects' products, whether it is code or documentation, to facilitate easier understanding of the algorithms and a straightforward code reuse capability for other interested parties. The standards enforcement itself is accomplished by appropriate team coordination and supervised by the EMD-MLR faculty. The latter of the SDP's component is the conduct of research by the teams utilizing their own software implementations and, potentially, third-party implementations of ML algorithms and/or approaches.

Despite the fact that the students spend a significant portion of their time to understand, code and document an existing ML algorithm, the ultimate goal is (a) to compare a number of ML algorithms or techniques on appropriately chosen benchmark problems, or (b) to apply the developed ML algorithm to solve a real-world problem that, either the supervising faculty has assigned to the student team. At the end of each year the project's Academic Committee evaluates the final products and research reports of each team and selects the best projects in terms of quality, which will then be submitted for publication to conferences and/or journals pertaining to education, ML or related areas.

As we mentioned earlier, SDPs are reminiscent of senior design projects for engineering students and capstone projects for computer sciences students (equivalent for two 3-credit hour courses, normally offered in the Fall and Spring of the student's senior year), which are well established educational vehicles and have become an integral part of engineering and computer science curricula. Owing to the proven educational effectiveness and value of these course projects, EMD-MLR's concept of SDP teams has been modeled after them. Community college students participating in these SDP teams are awarded two 1-credit hour coursework from the host University (FIT or UCF) that will be transferred towards their Associate degree at the corresponding community college (BCC or SCC).

Here we need to mention, that, before the formation of SDP teams takes place, EMD-MLR's offers to the body of participant students an introductory, month-long, ML course specifically designed to take into account the students' background and to address the necessary knowledge for the students to complete their projects.

Finally, the project features a strong assessment and evaluation plan. The development of all evaluation instruments is based on a variety of curriculum evaluation models. Tyler's Objectives Evaluation Model (Tyler, 1949), Stake's Contenance Model (Stake, 1967), and Eisner's Connoisseurship Model (Eisner, 1990). All these models are discussed in detail in the most recent books on curriculum development and evaluation (see March and Willis, 1999). Tyler's model is primarily used for the planning of the curriculum. On the other hand, Stake's model is utilized for the formative (on-going evaluation) of the curriculum, while Eisner's model is incorporated for the assessment and evaluation process (summative evaluation) at the completion of the EMD-MLR experience.

4. Project Merits

Overall, the proposed EMD-MLR program has many important features. First of all, it has a very important research component. Another important feature is that the teaching of the EMD-MLR students happens within a small group of undergraduate students of knowledge-diverse backgrounds. This group will work together towards a common goal (the completion of a research report). In this setting, the research motivates the teaching content. This framework is a model that encourages focus, coordination, cooperation, face-to-face interaction, interpersonal and small-group skills, and an eventual bonding among the students in the group, all of which are elements that enhance learning. The structure of the SDP teams and their associated goals shares a lot of commonalities with a cooperative learning environment. Many educational specialists have hailed cooperative learning as a very effective learning method compared to the more

traditional approaches of competitive and individualistic learning (see Johnson and Johnson, 1999). Occasionally, the structure of these SDP teams is multi-disciplinary, due to the nature of the intellectual focus area (machine learning). Multi-disciplinary groups of students working together towards the completion of a common goal provide real-life scenarios that are beneficial to the students involved.

Moreover, a third important feature of the EMD-MLR program pertains to the significant impact that this endeavor has on students. Every year the program involves 30 seniors and 10 community college students; in other words, it directly affects a total of 80 students during its lifespan. If this concept is appropriately disseminated to the affiliate Universities (as well as other Universities) or to disciplines other than the EE, CS, and CpE, the number of impacted students could increase quickly.

A final, key feature of EMD-MLR is the tremendous dissemination possibilities, primarily because of its very own nature. As we have mentioned earlier, it is our intention to establish a large scale, online repository. In a year's time alone it is feasible to have coded approximately 10 ML algorithms. If this effort continues at the same pace, 50 algorithms will be coded and experimented in five years' time.

6. Project Outcomes & Future Plans

Despite being less than one year old, the project has already impacted 31 undergraduate students originating from the 4 participating institutions. Due to their involvement in EMD-MLR, about three quarters of undergraduate participants are considering attending graduate school in the near future, while about half of the participating community college students are making plans to continue their education by pursuing a bachelor's degree in STEM disciplines. Also, 4 All-Hands Meetings were held so far, where participants across all 4 institutions had the opportunity to interact and present their ideas, their work and results to their peers and other invited guests.

Furthermore, 11 software implementations of well-known and novel Machine Learning algorithms and techniques were developed, most of which were neural based. As examples we mention the Probabilistic Neural Network, Growing Cell Structures and Fuzzy Adaptive System Ellipsoid ARTMAP among others. Almost all implementations are provided as source code and MATLAB MEX files. Moreover, the produced software is accompanied with demonstrations and relevant documentation describing the essence of each algorithm or technique, so that it can serve as educational material for both introductory/undergraduate and advanced graduate curricula, as well as for research purposes in Machine Learning.

The most noteworthy outcomes of the students' guided research within the project's framework were disseminated in early 2005 to an international journal (Neural Networks) and 4 international/regional conferences (International Joint Conference on Neural Networks, International Conference on Parallel Processing, ASEE Annual Conference & Exposition and the National Institute for Staff & Organizational Development Conference). In total, Project EMD-MLR has supported so far 2 journal papers, 5 conference papers and 1 invited talk.

Due to the broad scope and ambitious vision, our plan for the project's remaining duration encompasses a variety of action items, whose timely and/or successful accomplishment will play a crucial role in the project's overall impact. First of all, a total of approximately 50 students from the 4 institutions will be recruited to form the student participation base for the project's second year. Emphasis will be given into creating excitement about the project's research area. Moreover, more educational material is planned to be contributed in the form of software implementations to expand the already existing collection of ML algorithms and techniques. More specifically, we forecast 10 more implementations by the end of the project's second year.

Also, our guided research efforts with the new participants will be continued and our dissemination efforts to broad outlets such technical, educational journals and conferences needs to be intensified. A key dissemination component is the launching, update and maintenance of the project's Machine Learning Repository at www.machine-learning.org. The entire collection of software implementations will be disseminated through this outlet. In addition, repository's existence needs to be advertised via Internet newsgroups, news bulletins and other online outlets, such as Mathtools.com. Furthermore, we plan to initiate discussions and negotiations with the related industry for the adoption of the software implementations in real-life applications. Our Industry Advisory Board will play a crucial role in reaching this goal. Also, we need to take steps to create, update the ML introductory lecture notes and audio/visual material via the EMD-MLR website (my.fit.edu/EMD-MLR) and notify our academic affiliates of our Academic Advisory Board about the availability of this material, in order to aid in its dissemination at their institutions and elsewhere. Finally, we need to prepare and submit 3 chapters of an undergraduate Machine Learning book, based on the developed EMD-MLR materials, to ML experts to solicit their feedback.

7. Brief Summary

In this paper, we presented an overview of Project EMD-MLR, an NSF funded, proof-of-concept project that introduces undergraduate students to Machine Learning (ML) research through their involvement in supervised, student project teams. The main focus of the project is the development of educational material in the form of software implementations of state-of-the-art ML algorithms and accompanying documentation, as well as the dissemination of this material through an on-line outlet, namely a ML repository. Project EMD-MLR will impact 80 students in a span of 2 years, some of which are members of underrepresented groups. Furthermore, the project's focus enhances partnerships amongst 4 neighboring institutions, and many more affiliate Universities dispersed within the US and abroad.

References

[1] The Chronicle of Higher Education. The 1999-2000 Almanac. 1999.

[2] Darling-Hammond, L., Ancess, J., and Falk, B., Authentic Assessment in Action, N.Y.: Teachers College Press, 1995.

[3] Eisner, E. W., The educational Imagination, N.Y.: Macmillan, 3rd edition, 1990.

[4] March, C. J. and Willis, G., Curriculum: Alternative Approaches, Ongoing Issues, Columbus, Ohio, Merill, 1999.

[5] National Research Council. From Analysis to Action: Undergraduate Education in Science, Mathematics, Engineering and Technology. Washington DC: National Academy Press, 1996.

[6] National Science Board. Science and Engineering Indicators. Arlington, VA: National Science Foundation, 1998 (NSB 98-1).

[7] National Science Foundation. Science and Engineering Degrees by Race/Ethnicity of Recipients: 1989-96. Arlington, VA: National Science Foundation, 1999 (NSF 99-332).

[8] Project Kaleidoscope Volume I-What Works: Building Natural Sciences Communities, 1991.

[9] Stake, R. E., "The countenance of educational evaluation," Teachers College Record, 68, pp. 523-540, 1967.

[10] Tyler, R. W., Basic Principles of Curriculum and Instruction, Chicago, University of Chicago Press, 1949.

Acknowledgement

The authors would like to acknowledge the partial support from the National Science Foundation through a CCLI-EMD grant #0341601 entitled " PROJECT EMD-MLR: Educational Materials Development through the Integration of Machine Learning Research into Senior Design Projects ". Finally, the authors would also like to thank the two anonymous reviewers that provided useful comments and suggestions for this manuscript.

Biographical Information

GEORGIOS C. ANAGNOSTOPOULOS is an Assistant Professor in the Electrical & Computer Engineering department of Florida Institute of Technology. His research interests are statistical machine learning, neural networks and data mining.

MICHAEL GEORGIOPOULOS is a Professor of the Department of Electrical and Computer Engineering at the University of Central Florida. His research interests lie in the areas of neural networks and applications of neural networks in pattern recognition, image processing, smart antennas and data-mining. He is an Associate Editor of the IEEE Transactions on Neural Networks since 2001.

KEN PORTS is a Professor of the Department of Electrical and Computer Engineering at Florida Tech. He is also the Engineering Director of Florida TechStart, the university business accelerator. His interests include microelectronics, nanoelectronics and radiation effects, entrepreneurial behavior and culture, and business processes such as product to market, strategic planning and execution, and project management. Dr. Ports has 48 publications and 11 patents.

SAMUEL RICHIE is an Associate Professor of the Department of Electrical and Computer Engineering at the University of Central Florida. His research interests include surface wave device modeling, optical character recognition, video image processing, and biomedical instrumentation. He is serving as the Assistant Dean for Distributed Learning for the College of Engineering and Computer Science with operations including nine origination sites serving over 200 courses in engineering per year.

NELLY CARDINALE is an Associate Professor of Computer Science at the Brevard Community College. She teaches Unix System Fundamentals, Unix Networking, Principals of Programming, Java, Web Authoring, Web Programming Languages, Introduction to Computers and Micro-computers Applications. She is a member of the IEEE Computer Society and is a CompTia certified A+, Network+ and Linux+ professional.

MELINDA WHITE is a Professor at Seminole Community College in Sanford, FL. She teaches classes in computer programming and applications. Melinda has a Bachelor's degree in Mathematics and a Master's in Instructional Computing, both from the University of Florida. In addition to teaching for over 15 years, she has over 10 years of industry experience in Computer Programming and Systems Analysis.

VETON KËPUSKA is an Associate Professor of the Electrical and Computer Engineering Department at the Florida Institute of Technology. He has joined academia after over a decade of R&D work in high-tech Speech Recognition Industry in Boston area. His research interests lie in the areas of Speech Processing and Recognition, Speech Coding, Microphone Arrays, Neural Networks and Applications of Neural Networks in Pattern Recognition, Speech Processing and Recognition, Blind Source Separation, , Image Processing, Natural Language Understanding, Human Machine Interface. He holds a number of patents in speech recognition area.

PHILIP CHAN is an Associate Professor of Computer Science at Florida Institute of Technology. His main research interests include scalable adaptive methods, machine learning, data mining, distributed and parallel computing, and intelligent systems. He received his PhD, MS, and BS in Computer Science from Columbia University, Vanderbilt University, and Southwest Texas State University respectively. He is an Associate Editor for the Knowledge and Information Systems journal.

ANNIE WU is an Assistant Professor at the School of Electrical Engineering and Computer Science at the University of Central Florida. Her research interests are in the areas of genetic algorithms, machine learning, biological modeling, and visualization.

MARCELLA KYSILKA is a Professor and Assistant Chair of the Education Foundations Department at the University of Central Florida. She is active in her professional organizations and currently serves as Associate Editor of the "Journal of Curriculum and Supervision" (the scholarly journal of the Association for Supervision and Curriculum Development). Her research interests are in curriculum studies.