

AC 2009-1194: THE AMALTHEA REU PROGRAM: ACTIVITIES, EXPERIENCES & OUTCOMES OF A COLLABORATIVE SUMMER RESEARCH EXPERIENCE IN MACHINE LEARNING

Georgios Anagnostopoulos, Florida Institute of Technology

GEORGIOS C. ANAGNOSTOPOULOS is an Associate Professor in the Electrical & Computer Engineering department of Florida Institute of Technology in Melbourne, Florida. He is also the Director of the AMALTHEA REU Program. His research interests are statistical machine learning, neural networks and data mining.

Michael Georgiopoulos, University of Central Florida

MICHAEL GEORGIOPOULOS has received a Diploma in EE from the National Technical University of Athens, Greece, in 1981, and an MS in EE and a Ph.D. in EE from the University of Connecticut, Storrs, CT, in 1983 and 1986, respectively. He joined the University of Central Florida in 1986, where he is currently a Professor in the School of EECS. His research interests lie in the areas of Machine Learning and applications with special emphasis on ART neural networks. He has published his work in over 250 journal and conference venues. He has been an Associate Editor of the IEEE Transactions on Neural Networks from 2002 to 2006 and he is currently serving as an Associate Editor of the Neural Networks journal. He has served as the General Chair of the S+SSPR 2008 Workshops, a satellite event of ICPR 2008.

Veton Kepuska, Florida Tech

VETON Z. KĚPUSKA is an Associate Professor of the Electrical and Computer Engineering Department at the Florida Institute of Technology. He has joined the academia after over a decade of R&D work in the high-tech Speech Recognition Industry of the 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 the speech recognition area.

Kenneth Stanley, University of Central Florida

KENNETH O. STANLEY is an assistant professor in the School of Electrical Engineering and Computer Science at the University of Central Florida. He received a B.S.E. from the University of Pennsylvania in 1997 and received a Ph.D. in 2004 from the University of Texas at Austin. He is an inventor of the Neuroevolution of Augmenting Topologies (NEAT) and HyperNEAT algorithms for evolving complex artificial neural networks. His main research contributions are in neuroevolution (i.e. evolving neural networks), generative and developmental systems, coevolution, machine learning for video games, and interactive evolution. He has won best paper awards for his work on NEAT, NERO, NEAT Drummer, and HyperNEAT. He is an associate editor of IEEE Transactions on Computational Intelligence and AI in Games, the chair of the IEEE Task Force on Computational Intelligence and Video Games, and has chaired the Generative and Developmental Systems track at GECCO for the last three years.

Alison Morrison-Shetlar, University of Central Florida

ALISON MORRISON-SHETLAR is Vice Provost and Dean, Undergraduate Studies and Professor of Biology at the University of Central Florida. Her science research interests are in the molecular biology, biochemistry and physiology of Hydrogen-proton membrane transport proteins and her pedagogical research is in the area interactive teaching and learning strategies for any size classroom.

Pat Lancey, University of Central Florida

PATRICE M. LANCEY earned her B.A. from Brooklyn College, Brooklyn, New York, in 1974, and an M.A. and Ph.D. in Clinical Psychology from Wayne State University, Detroit, Michigan, in 1979 and 1996 respectively. She joined the University of Central Florida in 2001 where she serves as Director, of Operational Excellence and Assessment Support. Dr. Lancey coordinates the university wide Institutional Effectiveness Assessment process and supports assessment of academic programs and administrative departments. She also designs statistical studies to provide information about student engagement, institutional conditions that enhance student learning outcomes, progression, and retention to provide actionable reports to decision makers to include upper administration, faculty and staff. Dr. Lancey has served as the outside evaluator for several NSF funded grant projects. Prior to this, she held positions at Johns Hopkins School of Public Health, Westat, Inc., University of Alabama, and Palm Beach Community College. She presents papers and workshops for faculty and administrators on educational assessment topics at national and regional conferences and has acted as a consultant to other universities. Dr. Lancey serves as a reviewer for proposals in the area of assessment for the Association for Institutional Research and Southern Association for Institutional Research. She regularly works with faculty to develop research methodology and student learning assessment.

Paula Krist, University of Central Florida

PAULA S. KRIST is the Director of Assessment Support for the School of Leadership and Education Sciences at the University of San Diego. She works with all departments to support program and student learning outcomes assessment for faculty and staff. Previously the Director of Operational Excellence and Assessment Support at the University of Central Florida and the Director of Institutional Research and Assessment at Florida Institute of Technology, Dr. Krist regularly presents workshops on assessment topics and enjoys working with faculty on grant projects. Her Ph.D. in Educational Psychology is from the University of North Carolina at Chapel Hill.

Tace Crouse, University of Central Florida

TACE CROUSE is currently the Interim Director of the Faculty Center for Teaching and Learning at the University of Central Florida where she has served since 2004. From 1999-2004 she was on the faculty of the university's College of Education. From 1986-1998 she served in various positions at Brevard Community College including faculty member, department chair, dean, campus provost and Executive Vice President. Her B.S. degree in Mathematics (1972) and her M.S. in Mathematics Education (1974) were earned at the Shippensburg University of Pennsylvania. Her Ed.D. is in Educational Leadership (1993) from the University of Central Florida. Her research area is in the use of assessment to improve instruction.

The AMALTHEA REU Program: Activities, Experiences & Outcomes of a Collaborative Summer Research Experience in Machine Learning

Abstract

The AMALTHEA REU Program is a 10-week, summer research experience for science or engineering undergraduate students funded by the National Science Foundation since 2007 and featuring Machine Learning as its intellectual focus. Moreover, it is a joint effort of two collaborating universities in Central Florida, namely Florida Institute of Technology in Melbourne and University of Central Florida in Orlando.

Organizing, implementing and directing REU Sites is typically perceived as a demanding effort; while offering unique advantages, operating collaborative sites may impose an additional layer of challenges. In this paper our intention is to present the objectives of our program, its unique characteristics, and the structure and organization of our collaborative site. Furthermore, we would like to give an informative account of our activities across the various aspects of the program, such as marketing of the experience, recruiting of student participants, the summer experience itself and our dissemination efforts. Finally, we report on our outcomes accomplished so far, which include research products and evaluation results.

While our program is only entering into its third year of operation, we do hope that, by sharing our experiences and promising strategies to date, we will encourage and aid prospective REU Site directors to successfully plan for and operate collaborative sites.

1. Introduction

The AMALTHEA REU Program¹ is a collaborative effort between two closely-located universities, Florida Institute of Technology in Melbourne and University of Central Florida in Orlando. "AMALTHEA" stands for *Advances of Machine Learning in THEory & Applications*, which, in turn, stems from the program's full title "*REU Site: Collaborative Research: Advances of Machine Learning in Theory and Applications (AMALTHEA)*." Finally, REU stands for "*Research Experiences for Undergraduates*," which is the name of the National Science Foundation (NSF) program funding this effort. Overall, the project seeks to provide top quality educational experiences to a diverse community of learners through research participation in the area of Machine Learning (ML).

Machine Learning is nowadays a high-importance, ever-expanding discipline that draws concepts from a variety of fields, including artificial intelligence, cognitive sciences, information theory, statistics, mathematics, physics, philosophy and biology among others. On the other hand, automatic target recognition, earthquake prediction, gene expression discovery, intelligent credit fraud protection and affectionate computing, to mention just a few, are examples of cutting-edge applications of ML in various technological and scientific domains. The project's thrust area is the theory of ML and how it can be integrated and applied to important real-life problems, thus exposing participants to both theory and applications.

As mentioned earlier, the AMALTHEA effort is funded and supported under the NSF's REU program² which states that it "...supports active research participation by undergraduate students in any of the areas of research funded by the National Science Foundation" and constitutes one of the several NSF programs that aim to develop a diverse and globally-competitive workforce of future US engineers and scientists. Project Kaleidoscope (PKAL)³ an informal alliance of faculty, focuses on building learning environments that attract and sustain undergraduate students to the study of STEM (science, technology, engineering and math) fields and motivate them to consider careers in related fields. PKAL (funded in part by NSF and in which one of our educational specialist and principal investigator is a current assistant director), emphasizes that "...the undergraduate years are the last opportunity for academic study of STEM 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 significant step in the right direction. Additionally, based on national data^{4,5} the percentage of bachelor's degrees for under-represented 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. Therefore, the involvement of students from under-represented groups in programs such as the REU is deemed imperative.

The AMALTHEA REU Program takes part in this national effort by involving an average of 10 undergraduate students annually in ML research over 10 weeks in the summer (mid May through end of July). The 3-year project is support by NSF through grants No. 0647018 and No. 0647120 at \$160,701 and \$138,750 respectively. Apart from the cumulative NSF support of almost \$300,000, most of which goes to REU student stipends, housing and subsistence expenses, the host universities are contributing an additional \$90,000 to support REU and graduate student stipends. Overall, it plans to impact a diverse group of at least 30 students, as well as about 12 graduate students, which will participate in undergraduate teaching and mentoring activities. The undergraduate students perform supervised research on ML topics that have the potential to impact the field of ML itself, as well as how ML is applied to other scientific disciplines. REU research results are expected to be published in interdisciplinary conferences, and, potentially, in technical journals. Additionally, these REU research advances are fed back and integrated into the teaching of ML-related courses at the partnering institutions. AMALTHEA features two main objectives, which can be summarized as follows:

- To recruit a diverse, talented body of undergraduate students from around the nation and from a variety of engineering and science disciplines. Furthermore, the program is particularly interested in reaching women, minorities, and, in general, student groups that have had traditionally low representation in science, technology, engineering and mathematics disciplines. Toward this end, the program is supported by a geographically-broad, national network of Affiliate Universities, some of which serve significant numbers of underrepresented minorities and others offer only BS / BA degrees.

- To offer an experience that will actively engage the recruited students into cutting-edge Machine Learning research. The program aims to form, maintain and evolve a vibrant community of learners here in Central Florida, which will foster and provide a valuable summer research experience for undergraduate students through participation in research programs and high quality student/faculty interaction and mentorship. We plan to familiarize and excite the participant students about many, state-of-the-art aspects of ML, which, we hope, will facilitate their retention in STEM fields, either career-wise or by continuing into STEM graduate education.

Our Program is supported by a network of affiliate universities and faculty across the country. The project staff maintains a close collaboration with these faculty on matters such as student recruitment, formulation of research topics, dissemination of the projects' results and products, assessment and evaluation among other issues. Additionally, the AMALTHEA REU Program aims for developing and maintaining close ties to the local industry and government sectors. It is supported by a number of industry affiliates and collaborators that form the program's Advisory Board (AB). The board's role is to contribute their technical expertise and aid in the assessment and evaluation process of the entire program. More specifically, the AB members take part in AMALTHEA's Symposium at the end of the summer experience and help with the program's cumulative evaluation. To this end, they assess the technical quality of the research outcomes and the quality of the program as a whole by interviewing participants and gauging various other aspects of the experience.

The rest of the paper describes the various components and outcomes of our Program over the years 2007 and 2008. In particular, since REU sites that are being run collaboratively among 2 or more host universities are a rare phenomenon (at the time of writing the authors are aware of two more collaborative REU Sites funded by NSF's Directorate for Computer and Information Science and Engineering) we hope that by sharing our experiences and promising strategies to date, we will encourage and aid prospective REU Site directors to successfully plan for their own collaborative sites. Therefore, Section 2 of our paper discusses our marketing and recruiting activities prior to the actual summer experience. Section 3 provides a description of activities during the summer experience, while Section 4 discusses the outcomes achieved the past 2 years. Finally, Section 5 concludes with a discussion about our findings.

2. Marketing & Recruiting Activities

Advertising the offered program and then recruiting participants for it is one of the most important pre-summer activities of the AMALTHEA effort. The marketing and recruitment strategies that are being employed greatly determine the quantity, quality and make-up of the body of applicants that will apply to the Program and, therefore, their importance cannot be overstated. During the first year of the Program (2007) implementing an effective advertising campaign and an aggressive recruitment plan was especially challenging due to the very limited time (about a month) between the official award of the REU grant and the application submission deadline.

From the Program's beginning it became clear that a website was needed **(a)** whose URL could be communicated and advertised, especially via electronic means like mass emails and website links **(b)** that would provide sufficient detail to potential applicants about the Program's nature through FAQ pages, **(c)** that would enable on-line application submission and **(d)** that would facilitate the on-line review of applications and the final selection of participants. While satisfying criterion (c) greatly reduces the application management logistics and keeps application data organized in a database, meeting criterion (d) is quite important for collaborative REU sites, whose project staff members may be geographically distributed. Thus, a website was designed and developed with PHP-based backend functionality, so that application data would be stored in a MySQL database. The following year, project descriptions, posters and publications from 2007 were also hosted on the website to advertise our summer research experience. Additionally, the on-line application submission forms and the on-line application reviewing mechanisms were fine-tuned and enhanced, where it was necessary.

Regarding promotional material that would be suitable for mass dissemination and advertisement, the directors of the effort sought out professional expertise for designing them so as to provide AMALTHEA with a competitive edge among other REU site advertisements. Thus, the Office of Marketing of one of the host universities provided the Program with flyer, poster and pull-up display designs for both years. Emails and electronic versions of the flyer were mass-distributed to juniors and seniors of the Honor's College, Engineering College and the Chemistry, Physics, Mathematics, Statistics, Biological Sciences departments of both host universities. In addition, we advertised our REU site to students of local ACM, IEEE, HKN, Women in EE & CS (WEECS), Society of Women Engineers (SWE) and Women in Engineering (WIE) chapters. Furthermore, posters and pull-up displays were prominently displayed at the host universities.

Apart from the local advertisement at the host universities, in order to advertise at a national scale the project staff emailed softcopies and mailed hardcopies of the flyer to every member of the Electrical & Computer Engineering Department Head Association (ECEDHA; 339 addressees) as well as any department head or institution listed in the Computing Research Association's (CRA) Forsythe List (302 addressees). In other words, we mailed a total of about 640 flyers in both paper and electronic form for both years. Moreover, we resorted to our network of 10 Affiliate Faculty (some of which reside at 4-year colleges or institutions that serve a large number of student from under-represented groups) that helped us with distributing 100 flyers each at their home institutions and perform personal recruiting of potential applicants.

In 2008, we were able to take additional steps to secure a high-quality applicant pool. When attending the Annual Principal Investigators' meeting of all the CISE-supported REU Sites, which was held at Austin, Texas on February 28-29, 2008, we distributed posters, flyers and 2007 proceedings for the attendees, so that they can advertise our Program at their own institutions. In addition, the following national organizations were contacted via email to promote the Program to their members or affiliates:

- Association of Women in Mathematics
- Society for Industrial and Applied Mathematics (SIAM)
- American Mathematical Society (AMS)
- American Statistical Society (ASS)
- Society for Women Engineers (SWE)
- National Center for Women & Information Technology (NCWIT)
- Women in Engineering Organization
- Women in Engineering Programs \& Advocates Network (WEPAN)
- IEEE Women Society
- ASEE Student chapters
- Computing Research Association (CRA-W)
- Association for the Advancement of Artificial Intelligence (AAAI)
- National Society of Black Engineers (NSBE)
- Society of Hispanic Professional Engineers (SHPE)
- Anita Borg Institute (ABI) for Women and Technology
- American Computing Association (ACM-W)
- American Computing Association's (ACM) Special Interest Group in CS Education (SIGCSE)

By the submission deadline of March 30th, the AMALTHEA Site featured was able to attract 65 eligible applicants in 2007 and 62 in 2008. Of these applicants, 20 were phone-interviewed and a participation offer was eventually extended to the top candidates. In 2007, 10 offers made yielding 7 participants and in 2008 13 offers were made yielding 13 participants. Crude profile characteristics of our participants for these last 2 years are provided in Table 1 and Figure 1. The data demonstrate that 50% of the Program's participants over the past 2 years belong to groups that are typically under-represented in traditional STEM fields.

Table 1: Participant body characteristics for the last 2 years of the Program.

Participants	2007	2008	Total
Hispanic/Black Males	1	4	5
Hispanic/Black Females	1	0	1
Non-Hispanic/Black Males	4	6	10
Non-Hispanic/Black Females	1	3	4
Total	7	13	20

3. Summer Experience Activities

As mentioned in the Introduction, the AMALTHEA REU Program lasts 10 weeks and typically spans the period of mid-May through end of July every year. During the first day all REU students at each site participate in an orientation regarding the Program. Participants are introduced to the rest of the AMALTHEA staff, are familiarized with the university surroundings (eateries, laboratories, library, emergency contacts, etc.), complete a variety of participation forms, are given access to a variety of resources (such as email accounts, access to labs, etc.) and, finally, are introduced to the essential details of the Program's structure, schedule and activities.

A. The Machine Learning Primer Course

Immediately after Orientation, the Machine Learning Primer (MLP) Course is held for 4 days (Tuesday through Friday during week 1). During this time period the entire community meets at a common location (one of the host universities). The morning part of the schedule typically consists of a lecture session, followed by a communal lunch break, followed by a lab session in the afternoon, which is conducted by graduate students.

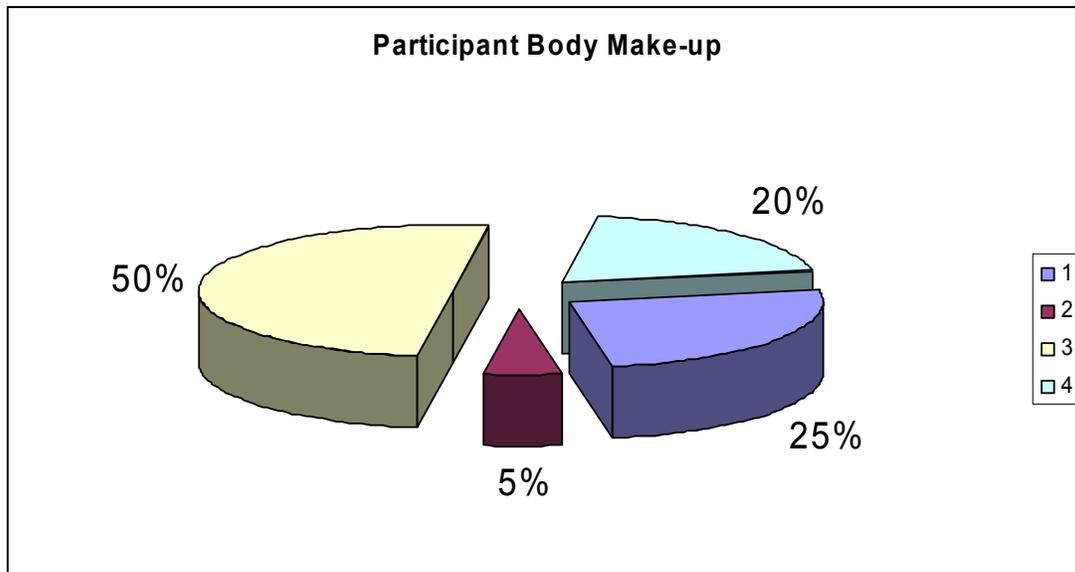


Figure 1: Cumulative percentages of different student groups that have participated in the Program over 2007 and 2008. Groups 1, 2, 3 and 4 correspond to Hispanic/Black Males, Hispanic/Black Females, Non-Hispanic/Black Males and Non-Hispanic/Black Females respectively.

The lecture sessions were organized and performed by the faculty mentors and their goal was to provide a necessary, rudimentary foundation in ML. Through these lectures the REU students were given the opportunity to familiarize themselves with our general research theme, as well as to get exposed to more special knowledge that helps them to begin from early on to explore their corresponding research projects. Lecture material included an overview of ML and its impact in everyday life (day 1), an introduction to pattern recognition (day 2), an overview of clustering (day 3) and an introduction to regression (day 4). In turn, each of these lectures attempts to emphasize and promote the understanding of important concepts that are relevant to the lecture's topic. For example, during day 2 specific concepts discussed were the Bayes decision criterion, various simple classification models (like k-Nearest Neighbor, Parzen Windows Classifier, Decision Trees, single-layer Perceptron, etc.), and model assessment and selection methods (such as cross-validation, etc.). Before the MLP Course a pre-knowledge questionnaire is administered. The results of this test are then correlated with the results of an identical post-knowledge questionnaire answered by the students after the MLP Course's completion. The comparison of the questionnaires provides a quantitative measure of student learning / understanding and, to some extent, of the course's effectiveness. Additionally, during the course, the participants are asked to provide daily feedback about the course (lecture & lab)

material and the instructor(s). The analysis and assessment of all these aforementioned results are performed by our Assessment & Evaluation (A&E) specialists. Apart from providing a basic knowledge platform to each student that is unfamiliar with ML, the MLP Course is also an instrument to bring all student participants together, get to know each other and to start developing a sense of community among them, which is crucial for other activities in the weeks to come, such as social activities.

Following the course, the student participants are introduced to their research topics and in the subsequent weeks continue to acquire more specific technical knowledge and further develop their expertise via close interactions with their graduate and faculty mentors throughout the summer experience.

B. Day-to-day Activities

Monday through Thursday the participant students at each site would attend a Briefing in the morning. During the meeting, mentors (graduate students and faculty) would meet with all REU students at the site and brief them on the activities and plans for the day. The briefing's goal was (i) to ensure that everyone in the teams had a clear understanding of what needed to be accomplished for the day, (ii) to ensure that all necessary implements were in place to materialize these plans, as well as (iii) to address any foreseeable problems. After the briefing research work would commence in the laboratories. Around noon the students and faculty would go out for lunch together as a group to promote team-building. Occasionally, time was provided to the students to run a few daily errands. Work would resume after lunch until later in the afternoon, when students would author and submit their journal entries. In 2007 each student emailed her/his journal entry to their corresponding faculty mentor, which would read it and meet in person with the student, if the situation demanded it. As of 2008 the authoring and archiving of daily journal entries is being done through Moodle⁶, which is an online content/course management system.

The weekday would conclude with a Debriefing, where all students and mentors would gather again to reflect on the activities and accomplishments of the day. During the Debriefing a short account was discussed of what had transpired during the day, what had been accomplished, what problems had occurred, how they were overcome, etc. Additionally, plans were forged and coordinated for the next day, if applicable. It was rather common that, after the daily Debriefing, the project staff from the two sites would communicate with each other about the day's progress and issues encountered.

C. Research Activities

Student participants in AMALTHEA are typically organized in research teams typically consisting of 1-3 REU students, a graduate mentor (when possible) and a faculty mentor. The research topic pursued is typical one of immediate interest to the graduate mentor and/or the faculty mentor. Each team is given access to computer equipment in the same room, where they all work side by side with their mentor(s) throughout the week. Part of Thursday afternoons is usually dedicated to preparing progress reports and presentations for next day's All-Hands

Meeting, where they present to the entire community and they attend technical talks (see next subsection).

By the end of the summer experience each research team is expected to produce a Technical Report (TR) describing their research topics and findings, slides for their final presentation during the Symposium (see subsection 3.F) and, possibly, a demonstration, if applicable. During the course of the summer experience, the mentors provide guidance to their mentees about matters of technical writing, ethics and scientific methodology. The write-up of the TR is being performed incrementally and a first draft is expected by week 7 of the summer experience. Eventually, the TRs and posters are collected in electronic form and posted on the Program's website.

D. All-Hands Meetings

Almost each Friday the entire AMALTHEA community meets for the day at a common site, alternating between the campuses of both host universities. These meetings are referred to as All-Hands Meetings (AHMs). In total, each year 7 of them are held (end of week 2 to week 8). The purpose of these AHMs was to (i) bring the geographically-distributed AMALTHEA participants together, (ii) create a sense of community and (iii) facilitate meaningful / useful interactions between the two sites through a series of common activities.

A typical AHM would start in the morning with student presentations regarding their research topics. During their presentation, student participants would provide an overview of their research topic, explain necessary concepts, state the goals and ramifications of their research, report progress and challenges and, finally, showcase intermediate or final results. The presenting students would then receive feedback by the rest of the community. After a communal lunch break, the AHM would usually continue with an invited talk or a special presentation by a project staff member, an outside faculty or an industry affiliate. These talks are usually of technical nature (that is, related to ML), while occasionally they are about preparation for graduate school, presentation skills, etc. Following the seminar, are given time to work on so-called special projects, which were, in essence, fun-activities and in-secret preparation of satirical material. This material, which was meant to satirize various aspects of the Program, was not revealed to the project staff until the end of the Symposium. The purpose of this activity was to further enhance the camaraderie among the student participants, to provide an alternative outlet for the students' creativity and to act as a pleasant diversion from the Program's routine.

At the end of the day student participants would author their journal entries and a communal Debriefing would follow. The AHM would often conclude with a social activity (e.g. a BBQ). While the students worked on their special projects, all graduate and faculty mentors would meet for a Project Staff Meeting (PSM). During the meeting, the staff would assess the research progress, discuss issues that occurred and solutions to overcome them and, finally, plan / coordinate future activities of the Program. It should be noted that a few PSMs in the beginning of the summer experience were held via teleconference during normal weekdays to immediately address some pressing logistics. Finally, the Mid-Project Focus Group takes place during the 4th AHM (end of 5th week). The focus group is performed by our Assessment & Evaluation (A&E) staff in absence of the project staff. Its purpose is to solicit candid opinions from the students

about the Program and the staff. A report is compiled a week later, which is then forwarded to the staff and discussed in order to consider corrective actions.

E. Social Activities

Social activities, being of communal nature in a setting outside of what is perceived as a “work environment”, are probably the single most important ingredient to establish cohesiveness among the participants. In AMALTHEA, social activities are voluntary and most of them are organized by the students with the help of the staff, which provides transportation and/or logistical support. The organizing and scheduling of such events creates an additional need for frequent interaction between the two populations of students hosted at different campuses.

F. The Symposium

The Symposium constitutes the apex of the AMALTHEA REU Program. It is held on the last Friday of the summer experience at one of the host universities is the final, formal activity of the Program, during which the entire AMALTHEA community meets with its AB and outside invitees to showcase its efforts and accomplishments.

The Symposium usually starts with an overview presentation of the year’s summer experience and the introduction of the student participants and project staff to the audience of AB members and Symposium invitees. The Symposium continues with the students’ oral presentations of their work and obtained results. Each presentation was followed by a brief Q&A session, during which the audience asked the students pertinent questions about their research topic. Also, for each presentation the AB members filled out an evaluation rubric to be utilized for feedback purposes. Furthermore, a poster session was held during the extended lunch break. Students attend their posters, while AB members and the other invitees (such as university dignitaries, etc.) have the opportunity go around, mingle with the participants, visit each poster and have a closer interaction with the students. When applicable, project teams also present live demonstrations related to their project.

After lunch and the poster session, the project staff and outside invitees remove themselves from the Symposium room and our A&E experts conduct the End-Project Focus Group in the presence of the student participants and the AB. During this focus group the students were asked their opinion on various aspects of the Program. The focus group discussion is guided by a special questionnaire, while additional questions are usually posed by the AB. After the completion of the focus group the students were excused and the project staff is invited back in to take part in an open discussion about the focus group’s major findings and to receive direct feedback from the AB. After the end of the aforementioned forum, the AB members and the A&E staff are thanked for the services and the Symposium is officially concluded. A brief special projects presentation session is held in private (only participant students and project staff are present), where the student participants satirize various aspects of the summer experience through audio-visual means. The last formal activity of the Program is a communal, festive dinner the same evening. Next day, out-of-town participants surrender their access means to their lodgings, finalize their business with the host universities and return to their place of origin.

4. Outcomes

A. Research Outcomes

In 2007 the following research topics were pursued:

- A Backward Adjusting Strategy for the C4.5 Decision Tree Classifier⁷
- A Grid Based System for Data Mining Using MapReduce⁸
- NEAT Drummer: Interactive Evolutionary Computation for Drum Pattern Generation⁹
- Testing and Improvement of the Triple Scoring Method for Applications of Wake-up Word Technology¹⁰

On the other hand, in 2008 these topics were researched:

- Real-time, Static and Dynamic Hand Gesture Recognition for Human-Computer Interaction¹¹
- Multi-stage Automatic License Plate Location & Recognition¹²
- Development of a Large Vocabulary Continuous Speech Recognition System for Rich Transcription Evaluation Using HTK¹³
- Detecting Outliers in Categorical Data Sets Using Non-Derivable Itemsets¹⁴
- Interactively Evolved Modular Neural Networks for Agent Control¹⁵
- Iterative Inner Solvers for Revised Simplex SVM Training¹⁶

Overall, the corresponding TRs and posters (a total of 10 each) were compiled and published on our website. Examples of a poster and TR are shown in Figure 2 and Figure 3 respectively. The interested reader is referred to these TRs and posters for more details on the particular scope and outcomes of each project. In a nutshell, since 2007 the AMALTHEA REU Program has directly supported 2 Honors in the Major theses of 2007 student participants^{17,18}, 1 Masters thesis¹⁹ and 1 doctoral dissertation²⁰ produced by former graduate mentors. Additionally, it has partially/indirectly supported an additional Masters thesis²¹. The REU research has also led to the publication with former undergraduate participants of 5 conference papers²²⁻²⁶ (with undergraduates Hoover, Beck, Garcia, Cardona, Alexander, Ahmed, Sparks and Miguez) and 1 journal paper²⁷ (with undergraduate Hoover). Also, it has indirectly and subsequently aided in the research presented in 9 additional conferences and 2 journals, which we do not list here.

B. A&E Outcomes

First of all, A&E instruments utilized in relation to the MLP Course were **(i)** the pre-knowledge questionnaire (direct measure) to assess the level of pertinent knowledge for each student, **(ii)** daily satisfaction surveys (indirect measures) that were used to evaluate the MLP instructor(s), the lectures and the labs and **(iii)** the post-knowledge questionnaire (another direct measure), which was used to assess individual knowledge gained from the course and, to an extent, the effectiveness of the MLP Course.



Detecting Outliers in Categorical Data Sets Using Non-Derivable Itemsets

Michelle Fox, Gary Gramajo, Anna Koufakou and Michael Georgiopoulos

MOTIVATION

- Traditional outlier detection unsuited to categorical datasets
- Current methods for categorical data use frequent itemsets (FI) with candidate FI growth exponential to number of attributes
- Real world datasets are very large, computationally expensive methods infeasible for real world applications
- Outlier detection has applications such as credit card fraud detection and network intrusion discovery

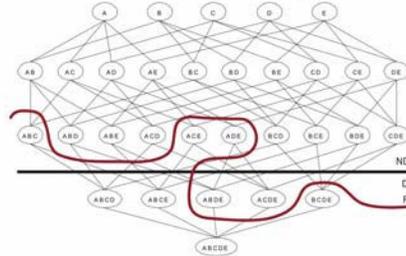
OBJECTIVE

Improve efficiency & scalability of outlier detection in categorical datasets using a condensed representation of frequent itemsets

METHOD

- Find frequent itemsets based on the Apriori and Non-Derivable Itemset (NDI) Algorithms
- For each representation of frequent itemsets, outlieriness of a datum is calculated using an appropriate scoring function
- Apriori-OD and NDI-OD are outlier detection strategies based on Apriori and NDI, respectively

Figure 1. Non-Derivable Itemsets, Derivable Itemsets and Frequent Itemsets



NDI Deduction Rule Generation:

$$\text{supp}(I) \leq \sum_{X \subseteq J \subset I} (-1)^{|J/I|} \text{supp}(J) \text{ if } |I/X| \text{ is odd}$$

$$\text{supp}(I) \geq \sum_{X \subseteq J \subset I} (-1)^{|J/I|} \text{supp}(J) \text{ if } |I/X| \text{ is even}$$

Figure 2. Calculating support interval for itemset I = {a,b,c}. Support interval of I = [2,2] (using deduction rules above), LB(I) = UB(I) = 2, I is derivable => itemset I pruned

tid	Items
1	a,b
2	a,b,c,d
3	a,b,c
4	a
5	c

- $R_{abc} : \text{supp}(I) \geq 0$
- $R_{ab} : \text{supp}(I) \leq \text{supp}(ab) = 3$
- $R_{ac} : \text{supp}(I) \leq \text{supp}(ac) = 2$
- $R_{bc} : \text{supp}(I) \leq \text{supp}(bc) = 2$
- $R_a : \text{supp}(I) \geq \text{supp}(ab) + \text{supp}(ac) - \text{supp}(a) = 1$
- $R_b : \text{supp}(I) \geq \text{supp}(ab) + \text{supp}(bc) - \text{supp}(b) = 2$
- $R_c : \text{supp}(I) \geq \text{supp}(ac) + \text{supp}(bc) - \text{supp}(c) = 1$
- $R_{\emptyset} : \text{supp}(I) \leq \text{supp}(ab) + \text{supp}(ac) + \text{supp}(bc) - \text{supp}(a) - \text{supp}(b) - \text{supp}(c) + \text{supp}(\emptyset) = 8$

RESULTS

Figure 3. Runtime performance (seconds) for NDI-OD vs. Apriori-OD for KDDCup 1999 set; for min support lower than 99% Apriori-OD is computationally infeasible

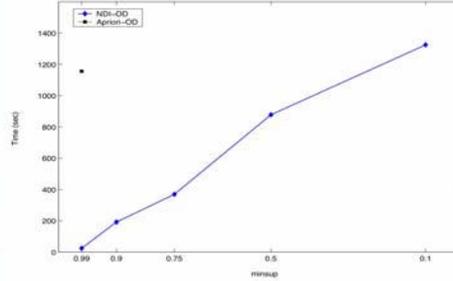


Table 1. Runtime performance (in seconds) and Generated Sets for NDI-OD and Apriori-OD for the KDDCup 1999 dataset; minsup: minimum frequency threshold

minsup	Runtime		Generated Sets	
	NDI	Apriori	NDI	Apriori
99.4%	24	1156	177	6349
90%	192	—	1513	> 85000
75%	369	—	2845	—
50%	878	—	7262	—
10%	1325	—	16818	—

Table 2. Correct Detection (CD) and False Alarm (FA) for NDI-OD versus Apriori-OD for the KDD 1999 dataset (minsup = 99.4%); k: target number of outliers

k	Correct Detection		False Alarm	
	NDI	Apriori	NDI	Apriori
1309	0.53	0.65	0.0092	0.0075
1500	0.64	0.65	0.0109	0.0095
1720	0.64	0.65	0.0118	0.0117
2000	0.70	0.70	0.0141	0.0141

DATABASES

- Breast Cancer — labeled malignant and benign, 699 points, 9 attributes, every 6th malignant record kept resulting in 444 non-outliers (malignant) and 39 outliers (benign)
- KDDCup 1999 — network intrusions on military computers, continuous attributes discretized, processed dataset contained 98,587 points, 39 attributes, 1,309 attacks/intrusions

CONCLUSION & FURTHER RESEARCH

- NDI-based outlier detection is significantly more efficient & scalable than Apriori-based outlier detection with comparable accuracy
- Further research to include applying NDI-OD to other datasets and comparing performance against Apriori-OD
- Also, implementing NDI-OD for data that are hosted at geographically distributed sites



This material is based upon work/research supported in part by the National Science Foundation under Grant No. 0647120 and Grant No. 0647018



Figure 2: Sample project from the 2008 AMALTHEA REU Program that represents the work of undergraduates Michelle Fox and Gary Gramajo with graduate mentor Anna Koufakou and faculty mentor Prof. Michael Georgiopoulos on using the Non-Derivable Itemsets representation for outlier detection on categorical data. The outcomes of their data mining related research have been submitted for publication and are pending review.

Detecting Outliers in Categorical Data Sets Using Non-Derivable Itemsets

Michelle Fox^a, Gary Gramajo^b,
Anna Koufakou^c and Michael Georgiopoulos^d

^aSchool of EECS, Milwaukee School of Engineering, Milwaukee, WI;

^bSchool of Mathematics, Florida State University, Tallahassee, FL;

^cSchool of EECS, University of Central Florida, Orlando, FL;

^dSchool of EECS, University of Central Florida, Orlando, FL

ABSTRACT

Outlier Detection is a research field with many applications, such as detecting credit card fraud or network intrusions. Most previous research focused on numerical data and pair-wise distances among data points to detect outliers. Nevertheless, most categorical data sets lack straightforward mapping to numerical values and approaches that rely on computing distances do not apply so easily. Recently, a few outlier methods were proposed for categorical datasets using the concept of Frequent Itemsets (FIs). The number of generated FIs can be far too high, especially in the case of large, dense datasets, containing a high number of categorical values. There has been much research towards summarizing and/or condensing the FIs in a dataset to address this issue. However these ideas have not been applied directly to the field of outlier detection. In this report, we explore the effect of using a condensed representation of Frequent Itemsets, called Non-Derivable Itemsets (NDI), on the accuracy and efficiency of an outlier detection method. Our experimental results indicate that NDI-based Outlier Detection offers significant gains in terms of speed and scalability over a FI-based outlier detection, while maintaining comparable detection accuracy.

Keywords: Outlier Detection, Categorical Data, Frequent Itemset Mining, Non-Derivable Itemsets, Condensed Representations

1. INTRODUCTION

Outlier detection is a research field with many applications, such as credit card fraud detection,¹ or discovering network intrusion.² In contrast to the aim of traditional data mining (i.e. to mine common or frequent patterns in the dataset), outlier detection approaches focus on detecting patterns that occur infrequently in the dataset.³ One of the most widely accepted definitions of an outlier pattern is provided by Hawkins:

An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.⁴

The majority of the existing research efforts in outlier detection have focused on datasets with a specific attribute type, mainly assuming that attributes are only numerical and/or ordinal. In the case of data with categorical attributes, techniques which assume numerical data need to first map the categorical values to numerical values, a task which is not a straightforward process (e.g., the mapping of a marital status attribute (married or single) to a numerical attribute).⁵ A second issue is that many applications for mining outliers require the mining of very large datasets (e.g. terabyte-scale data⁶). This leads to the need for outlier detection algorithms which must scale well with the size and dimensionality of the dataset.⁷ Data may also be scattered across various geographical areas, which implies that transferring data to a central location and then detecting outliers is impractical, due to the size of the data, as well as data ownership and control issues. Thus, the

Further author information: (Send correspondence to M.G.):

A.K.: E-mail: akoufako@mail.ucf.edu; M.G.: E-mail: mighaelg@mail.ucf.edu, Tel.: 1 407 823 5338

Figure 3: Sample TR from the same project as in Figure 2. All TRs in 2008 were authored in LaTeX to ensure a uniform look and follow a conference paper format. Prior to authoring, the REU participants were trained in the basic use of LaTeX as an electronic typesetting and document preparation system.

For both years, the correlation between pre- and post-knowledge results clearly indicated the success of the MLP Course to provide the student participants with the necessary, fundamental background material. The satisfaction survey results for each of the MLP Course's 4

days showed that the students expressed great satisfaction with the lectures and labs. Both were reported to be well organized, understandable, useful, and the instructors showed concern for the students' learning.

As mentioned in an earlier section, A&E instruments that were employed on a daily basis were the individual journal entries, which each student participant would author before the day's Debriefing. In the journal entry each student would reflect on what she/he had accomplished and learned that day along with the challenges she/he faced and how she/he planned to overcome them. The survey of the entries revealed that, overall, the student participants appreciated all that they had learned. Challenges mentioned were among other minor issues, like presenting in front of an audience and programming or debugging. Finally, it consistently appears as if all the students became more self-reliant with time and felt more confident with what they were doing by the end of the Program.

Both years, the Mid-Project Focus Group took place during the end of week 5 and revealed that, overall, student participants in AMALTHEA were very positive about their experience. For example, the 2007 group of students compared their experience to that of friends in other locations, who were involved in programs similar in nature to AMALTHEA and they indicated that theirs was a superior experience overall. In specific, they felt they were more engaged directly in research than were their friends.

Regarding mentorship, students were happy with the graduate student mentors. They felt the graduate mentors were a good conduit of information to the faculty members. Additionally, all students were complimentary about their faculty mentors. They felt that faculty mentors were genuinely interested in their experience, always willing to help work out obstacles, whether within the research or in their extracurricular lives.

Regarding the project conduct, AMALTHEA students seemed to unanimously like the structure and content of the Program and indicated that the week-long orientation and the MLP Course was an excellent way to begin this kind of research experience. Moreover, the invited talks by researchers in ML were always perceived as a beneficial dimension of the Program.

The End-Project Focus Group, which was held during the Symposium, both years, reflected the same positive image as the one from the Mid-Project Focus Group reports. For most participants, this was the first time they had the opportunity to go this in-depth with research and to work so closely with faculty on research projects. All students felt a sense of accomplishment and success regarding their research at the Program's conclusion, while several felt that they had helped exploring the topic and move the related research a bit forward. Simultaneously, AMALTHEA students also recognized that the 10-week lifespan of the summer experience is a short time to make larger impacts. There was a consensus among the participants that the Program helped their understanding of graduate research and what it entails. While students were not certain whether or not they will pursue ML research in the near future, a few considered it as a realistic possibility after the end of the Program. Student participants stated that they liked the following two aspects of the experience best: (i) the actual research and (ii) the opportunity to work closely with faculty members. Finally, it was mentioned that all students would recommend the AMALTHEA Program to their peers.

5. Discussion

From the participants' perspective, based on the A&E findings of the previous section it becomes apparent that major contributing factors to the high satisfaction rate of the AMALTHEA participants were the actual immersion into their research and the opportunity to work very closely with graduate and faculty mentors, who exhibited genuine interest in their progress and development through this experience. The pedagogy of students being able to perform research alongside with faculty mentors and graduate students allows the students to be in a circumstance, where they can observe their mentor at work, how (s)he thinks, how (s)he overcomes problems, and seems to be very effective on at least two levels; first, students feel more directly involved in the "real" operations of the faculty's research; second, it creates an energy and enthusiasm among the students under the faculty's guidance that motivates them to be more curious, work harder, and feel their experience is more valuable.

In conjunction with the aforementioned observations, a deeper element surfaces as an important reason to the success in matters of research activities of the AMALTHEA REU Program: the fact that, apart from the student participants, graduate mentors and faculty were equal stakeholders in the research products of the effort. On balance, it seems that vested interest of all parties involved in a team research setting is an important ingredient in meeting the team's research goals with success.

From the organizers' perspective, operating a successful REU site and providing a valuable, productive summer experience with a very positive and reinforcing effect on its participants is a rather challenging task, but at the same time it can be extremely rewarding to all who participate, including the organizing faculty and their institutions. The challenges lie in many factors, such as the immense logistical issues that need to be successfully addressed within the proper timeframe before, during and after the summer experience. The collaborative nature of REU sites, such as AMALTHEA, adds an extra dimension of complexity to the many tasks that need to be planned, implemented, tracked, assessed and readjusted. Nevertheless, we still believe that a project, again, such as AMALTHEA, with sufficient support, organization and infrastructure could be potentially scalable to incorporate additional organizations and/or institutions, as well as to include additional REU students in its summer research.

The organizers of the AMALTHEA REU Program feel that a big part of the success is due to the appropriate infrastructure and procedures that supported the operation of this collaborative effort. Communal components such as the introductory course in the beginning of the experience, the AHMs and the weekly social activities are catalysts in forming a distributed, yet cohesive community of mentors and mentees that prospers in an atmosphere of camaraderie. Especially during 2008, we also realized the importance of utilizing a course management system, such as Moodle, which supported AMALTHEA communication and planning needs through its on-line forum, messaging, calendar, daily schedule, blogging and many other features. Yet, the factor contributing, perhaps, the most to the successful operation and management of the AMALTHEA summer experience is the determination, dedication and diligence not only of the immediate project staff (graduate mentors and faculty), but also the valuable help of several facilitators from the host institutions (like administrative staff) and the surrounding industry (like our AB members), that invested their time to aid in materializing AMALTHEA's goals.

With respect to immediate future plans, the AMALTHEA REU Program will continue its course by offering another summer experience in 2009. Given the operating financial needs of such a project, REU programs are only able to sustain themselves with strong and multifarious institutional support, as well as with continuation of external funding. Therefore, its principal investigators plan to pursue additional funds from NSF and the private sector in 2009 to continue the Program for another 3 years.

Acknowledgments

This material is based upon work/research supported in part by the National Science Foundation under Grant No. 0647018 and Grant No. 0647120. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Finally, the authors are indebted to the four anonymous reviewers that provided suggestions to significantly improve our manuscript.

Bibliography

1. The AMALTHEA REU Program website, <http://cygnus.fit.edu/amalthea>, accessed on Feb 1st, 2009.
2. The Research Experiences for Undergraduates Program Page, National Science Foundation, http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5517&from=fund, accessed on Feb 1st, 2009.
3. Project Kaleidoscope Volume I-What Works: Building Natural Sciences Communities, 1991.
4. National Science Board. Science and Engineering Indicators. Arlington, VA: National Science Foundation, 1998 (NSB 98-1).
5. National Science Foundation. Science and Engineering Degrees by Race/Ethnicity of Recipients: 1989-96. Arlington, VA: National Science Foundation, 1999 (NSF 99-332).
6. Moodle. A virtual learning environment, <http://www.moodle.org>, accessed on Feb 1st, 2009.
7. Beck, J.R., Garcia, M.E., Zhong, M. Georgiopoulos, M., and Anagnostopoulos G.C. (2007) A Backward Adjusting Strategy for the C4.5 Decision Tree Classifier, Technical Report TR-2007-01, The AMALTHEA REU Program, Summer 2007.
8. Cardona, K., Secretan, J., Georgiopoulos, M. and Anagnostopoulos G.C. (2007) A Grid Based System for Data Mining Using MapReduce, Technical Report TR-2007-02, The AMALTHEA REU Program, Summer 2007.
9. Hoover, A.K., and Stanley, K.O. (2007) NEAT Drummer: Interactive Evolutionary Computation for Drum Pattern Generation, Technical Report TR-2007-03, The AMALTHEA REU Program, Summer 2007.
10. Stiles, A., Schmitt, B., Gertz, F., Klein, T., and Kepuska, V. (2007) Testing and Improvement of the Triple Scoring Method for Applications of Wake-up Word Technology, Technical Report TR-2007-04, The AMALTHEA REU Program, Summer 2007.
11. Hassan Ahmed, S.M., Alexander, T.C. & Anagnostopoulos, G.C. (2008). Real-time, Static and Dynamic Hand Gesture Recognition for Human-Computer Interaction, Technical Report TR-2008-01, The AMALTHEA REU Program, Summer 2008.
12. Li, R. Yassin-Fort, M. & Anagnostopoulos, G.C. (2008). Multi-stage Automatic License Plate Location & Recognition, Technical Report TR-2008-02, The AMALTHEA REU Program, Summer 2008.
13. Wax, D.A., Larsen, N.A., Furstoss, M.J. & Kepuska, V. (2008). Development of a Large Vocabulary Continuous Speech Recognition System for Rich Transcription Evaluation Using HTK, Technical Report TR-2008-03, The AMALTHEA REU Program, Summer 2008.
14. Fox, M., Gramajo, G., Koufakou, A. & Georgiopoulos, M. (2008). Detecting Outliers in Categorical Data Sets Using Non-Derivable Itemsets, Technical Report TR-2008-04, The AMALTHEA REU Program, Summer 2008.
15. Sparks, J.C., Miguez, R., Reeder, J. & Georgiopoulos, M. (2008). Interactively Evolved Modular Neural Networks for Agent Control, Technical Report TR-2008-05, The AMALTHEA REU Program, Summer 2008.

16. Astor, E.P., Lung, W.J. Ramirez-Padron, R., Sentelle, C.G. & Georgiopoulos, M. (2008). Iterative Inner Solvers for Revised Simplex SVM Training, Technical Report TR-2008-06, The AMALTHEA REU Program, Summer 2008.
17. Hoover, A.K., (2008). "Interactively Evolving Drum Tracks with Neural Networks," Honors in the Major Thesis, School of Electrical Engineering & Computer Science, University of Central Florida, Orlando, Florida, USA, Spring 2008.
18. Beck, J.R. (2007). "Implementation and Experimentation with C4.5 Decision Trees," Honors in the Major Thesis, School of Electrical Engineering & Computer Science, University of Central Florida, Orlando, Florida, USA, Fall 2007.
19. Klein, T.B. (2007). "Triple scoring of hidden Markov models in wake-up word speech recognition," Master's Thesis, Department of Electrical & Computer Engineering, Florida Institute of Technology, Melbourne, Florida, USA, Summer 2007.
20. Zhong, M. (2007). "An analysis of misclassification rates for decision trees," Doctoral Dissertation, School of Electrical Engineering & Computer Science, University of Central Florida, Orlando, Florida, USA, Summer 2007.
21. Nicoli, L.P. (2007). "Automatic Target Recognition of Synthetic Aperture Radar Images using Elliptical Fourier Descriptors," Master's Thesis, Department of Electrical & Computer Engineering, Florida Institute of Technology, Melbourne, Florida, USA, Summer 2007.
22. Hoover, A.K., Rosario, P.M., and Stanley, K.O. (2008) "Scaffolding for Interactively Evolving Novel Drum Tracks for Existing Songs," Proceedings of the 6th European Workshop on Evolutionary and Biologically Inspired Music, Sound, Art and Design (EvoMUSART 2008), in Applications of Evolutionary Computing, Mario Giacobini et al., Eds., Vol. 4974, (pp. 412-422), Napoli, Italy, March 26-28, 2008.
23. Beck, J., Garcia, M., Zhong, M., Georgiopoulos, M., Anagnostopoulos, G.C. (2008). "A Backward Adjusting Strategy and Optimization of the C4.5 Parameters to Improve C4.5's Performance", Proceedings of the 21st International Florida Artificial Intelligence Research Society (FLAIRS) Conference (FLAIRS 2008), David Wilson and H. Chad Lane, Eds., (pp. 35-40), Coconut Grove, Florida, USA, May 15-17, 2008.
24. Koufakou, A., Secretan, J., Reeder, J., Cardona, K., Georgiopoulos, M. (2008). "Fast Parallel Outlier Detection for Categorical Datasets using MapReduce", to appear in the Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN 2008), part of the 5th World Congress on Computational Intelligence (WCCI 2008), Hong Kong, China, June 1-6, 2008.
25. Alexander, T.C., Ahmed, S.M.H. and Anagnostopoulos, G.C. (2009). "An Open Source Framework for Real-time, Incremental, Static and Dynamic Hand Gesture Learning and Recognition," to appear in the Proceedings of the International Conference on Human-Computer Interaction (HCI 2009), San Diego, CA, July 19-24, 2009.
26. Reeder, J., Miguez, R., Sparks, J., Georgiopoulos, M., and Anagnostopoulos, G. (2008). "Interactively Evolved Modular Neural Networks for Game Agent Control," to appear in the IEEE Symposium on Computational Intelligence and Games (CIG 08), Perth, Australia, December 15-18, 2008.
27. Amy K. Hoover and Kenneth O. Stanley (2009). Exploiting Functional Relationships in Musical Composition, to appear in Connection Science, Special Issue on Music, Brain, & Cognition.