Jacek M. Zurada, Maciej A. Mazurowski, Rommohan Ragade,
and Artur Abdullin, *University of Louisville, USA*
Janusz Wojtudiak and James Gentle, *George Mason University, USA*

Research Frontier

Building Virtual Community In Computational Intelligence and Machine Learning

1. Introduction

It has been almost forgotten that not so long ago all research paper submissions were made in hard copy. At that time communication between far-away scientists could take months or years, thus making collaborations between distant teams slow and impractical. In the modern scientific community it is hard to imagine collaboration without e-mail, internet search engines, online submission sites, and other similar tools. Building on these modern conveniences, an even newer concept has arisen, namely of virtual organization.

A virtual organization is a group of geographically distributed individuals or institutions that cooperate with each other concurrently. With the advent of information technologies to help facilitate them, virtual organizations have not only become feasible, but also more convenient than traditional forms of organizations. This has made virtual organizations very popular in recent years. This paper tackles the issue of building a large-scale virtual organization for individuals and institutions that are associated with the field of computational intelligence (CI) and machine learning (ML). We begin with a few scenarios that will help illustrate the need for a virtual community in CIML.

As a first example, consider that many of us are researchers in CI and ML, and therefore are frequently reviewing papers for journals and conferences. It is quite usual that the reviewed papers are not within our primary line of research and we want to familiarize ourselves with the presented methods, results, and anything else that might be relevant. We may not find adequate information in the reviewed paper and its references. A more curious reviewer or reader may even want to find an implementation of the method and try it out, which is typically very difficult if at all possible.

Further, we might try to explore some application areas that are outside of our area of expertise. At other occasions we may want to undertake an

effort of merging or comparing multiple methods for one application. In such situations we typically need to contact other researchers, quickly obtain relevant reference material, or find programs that are easily adaptable to the project (possibly compatible with ours, and other already existing software). Each of these steps can impose tremendous challenges.

Another common scenario is that we develop a new method for a problem at hand. A natural next step is to compare this method to existing methods using available benchmarks. However, such comparisons may involve implementation of those existing methods, which is often a long and tedious process. It results in slowing down research or the avoidance of such comparisons altogether.

Finally, sometimes a researcher from another field, say chemistry, meteorology or medicine, a high school or college student, a person working for industry, or a medical doctor who wants to analyze data or solve a problem that is difficult or impossible to tackle with traditional methods. Locating and adapting CIML tools to solve such a problem may be very difficult for someone outside of the CI and ML field.

The above scenarios are obviously not rare or unique to CIML. In fact in many disciplines, particularly medicine, genomics, earth sciences, and some engineering areas, virtual communities are already well established to facilitate research and alleviate the difficulties mentioned above[1,2]. Long established initiatives like the National Library of Medicine (www.nlm.nih.gov), which hosts PubMed and Medline, includes almost anything that any professional or private person would want to search for in medical knowledge. Biologists and biotechnologists have their sources of data and publications at the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov) and Biomedical Informatics Research Network (BIRN; www.birn.net), which are further supplemented by multiple commercial websites. Jointly, this virtual community provides data (mostly genomics), computational tools,

Digital Object Identifier 10.1109/MCI.2008.930986

publication search engines, and everything else that researchers in the field may need.

Engineering, earth sciences, climatology, and similar areas' needs forced them to create virtual organizations whose main purpose is to share computational resources[3]. The use of grid computing[4] allows them to solve very large problems that are too computationally expensive to be solved at one location. In fact, currently the most notable efforts to create collaborative networks are based on the idea of grid computing, including the National Science Foundation supported TeraGrid project (www.teragrid.org) used across disciplines[5,6], and several domain-oriented projects such as MedIGrid[7].

It can be noticed that related virtual community building efforts in CIML are lagging in comparison with other domains. Although several attempts were made to create collaboration websites, data and software repositories, and also virtual communities, these efforts were not sufficiently synergistic and only addressed a small portion of the total CIML community needs. Among the most noticeable efforts in building CIML virtual communities are the PASCAL and PASCAL2 networks (www.pascal-network.org) which are European initiatives that support collaboration and research in cognitive systems. Areas of interest of both networks include machine learning, pattern analysis, machine vision, and natural language processing. Although the networks' website is rich in content (e.g. publications, video lectures, competitions), many items are only available to its members.

Most of CIML resources are distributed over dozens of websites maintained by individual researchers, groups, laboratories, and departments. These websites usually focus on specific topics of interest, and further, they lack any objective evaluations of the content on the site. Probably the most well known of these initiatives is an effort to implement popular machine learning algorithms in Java™ in the Weka system[8]. The software is available from the University of Waikato website www.cs.waikato.ac.nz/~ml/. In addition, some benchmark data can be found on a very popular site: UCI machine learning repository (http://archive.ics.uci.edu/ml/). Another resource is the evolutionary computation benchmarking site available at http://www.cs.ac.uk/research/projects/ecb/[9].

Wikipedia (www.wikipedia.org) is also a great resource for information. It follows the ideas of Web 2.0, in which users create content for the web. This content is, however, of varying quality that depends on the area, often incomplete, and sometimes simplistic or, on the other hand, tends to be too advanced for less experienced readers. Each article can be modified by multiple authors, and therefore it reflects their diverse knowledge. The downside of using Wikipedia is that its information often remains unverified. In contrast, papers published in archival journals most often undergo a very strict review which significantly increases level of confidence in the published material.

It is easy to see that these attempts, even though useful, are characterized by their concentration on selected aspects of virtual organization such as sharing software or data. We believe that integrating and unifying these aspects will be most beneficial. Such a broad and all encompassing structure for the CIML virtual community would provide not only access to its particular components or functions (such as data sharing or networking), but would also create a vast interconnection between the components, making such a community more integrated, better informed and poised for growth. In this paper we propose a framework for a CIML virtual community, discuss the issues crucial for its success, and present our initial efforts towards implementing it[10].

2. Building a Computational Intelligence and Machine Learning Community

Our goal is to create a virtual community that will eventually become a place for researchers, students, and the general public to find information about computational intelligence, machine learning, and related topics. Currently, we are pointing to five key aspects of virtual community in CI and ML. These are sharing data, sharing software, sharing computational resources, education and networking. We are planning to gradually include these components in the development of the CIML Community Portal. In this section we describe these aspects, their benefits, and their associated difficulties. We conclude this section with a discussion of, in our opinion, the most important feature of an effective virtual community in CIML, namely integration of all these components.

2.1 Sharing data

Sharing data can be one of the most important aspects of virtual communities in modern sciences and engineering. After developing a new algorithm, we are often interested in testing it. First we use simple, easily accessible benchmark problems. Then we use more complex state-of-the-art benchmarks. However, being engineers, the ultimate goal of our research is to solve real-life problems. Such problems typically belong to disciplines that are beyond our areas of expertise and therefore require data that is not immediately accessible, or is otherwise difficult to obtain.

Therefore, providing an accessible site for such data would be extremely useful. Despite its benefits, this, too, has associated difficulties. There are multiple restrictions on publishing data, especially when it involves human subjects (as in medical imaging). One needs to be very careful not to violate others' privacy while still providing useful information. Further, from the perspective of a CI and ML researcher, there exists an issue of formatting the data. This happens often because various types of software function only with data formatted in a certain way. This, in turn, often renders data formats incompatible with various types of software.

While at the moment we can not resolve the issue of restrictions on publishing certain proprietary real life data (other than encouraging the owners of the data to make it available to the public), we are planning to facilitate such publishing and to address the issue of

compatibility of data and software. In order to achieve it we will design a set of tools for publishing and translating data between popular formats, as well as encourage a common format.

2.2 Sharing software

Finding good and useful software data is difficult as is finding real life test or benchmarking data. There are many situations when we need software implementing certain methods or algorithms, or just their small components that could be incorporated in our own larger software. Sometimes simple tools are needed to just manipulate the data. Having easy access to a repository of high-quality CIML software could address these needs and would facilitate more efficient simulation-driven research or/and make it more thorough.

Similarly as with sharing data, there are some difficulties associated with sharing software. The most important, in our opinion, is that different pieces of software can be incompatible with each other. It is well known that during the time personal computers were commonly used for research in CI and ML, various languages (and various versions of them) were used to implement the developed algorithms. It means that many of these programs can not work together. This problem is very difficult to solve. To help alleviate it we encourage the use of Java to write and document software. Object oriented techniques used in Java allow for a very efficient integration of different programs and program components. At the same time Java can be integrated with other programming environments such as MATLAB, which is currently very popular in CI and ML research. We are also working on a framework allowing for better integration of existing software components that were written in other programming languages. Furthermore, the proposed framework will impose certain formats (or a list of allowable formats) for documentation, allowing for easier understanding of the functionality of software components.

Another difficulty associated with publishing software is that in most software sharing services everybody can post their programs without prior evaluation. An example of a popular site like this is sourceforge (http://sourceforge.net). We require for each submitted program, including its documentation, to be peer reviewed in a similar fashion as submissions to journals or reviewed conferences. Upon feedback from reviewers, authors may be able to improve their submissions before their final acceptance. This procedure ensures that only software that has met certain minimum quality requirements will be disseminated via the virtual community portal.

2.3 Sharing computational resources

Modern computational intelligence and machine learning algorithms are often computationally expensive, especially when applied to large real life problems. Therefore, running them on a single machine is often impractical. To alleviate this problem, many institutions share their computing resources. We are planning to develop a framework that will allow the sharing of resources on a larger scale. This will in turn allow CI and ML researchers to perform more comprehensive and faster experimentation. Developing such a framework will mean that several difficulties must be overcome, such as compatibility of different platforms, security issues, etc. Therefore, this functionality of the portal is planned for later stages of CIML virtual community building.

2.4 Education

Computational intelligence and machine learning are very dynamic disciplines with thousands of people involved in research and development of new theories, methods, and implementations. To facilitate effective research, easily accessible information about new discoveries is crucial. Certainly, there are multiple highly-rank journals and conferences that serve as outlets for announcements of new discoveries. Those media, however, are typically addressed to experienced researchers. Introductory and comprehensive material for new methods is often difficult to find.

We propose a platform that will aid in the sharing of educational materials intended for individuals new to the field such as students, or experts in other disciplines wanting to use some modern CI and ML methods in their fields. In addition, these materials (overview papers, tutorials with example programs) will hopefully be helpful for teachers at different levels while preparing their courses. In the future we also plan to extend educational materials by providing links to recorded lectures hosted by such services as videolectures.net or YouTube (www.youtube.com). There is no need to duplicate the efforts in recording them and hosting – as particularly videolectures.net provides access to many high quality lectures by well known researchers from CIML field. Our intention is to connect existing lectures to articles and other educational material in the CIML community portal.

2.5 Networking

Sharing high quality resources and providing educational materials is only part of the virtual community. The community is formed by the people and for the people – all those who are interested in CIML. One of the most important aspects of such a virtual community is providing an opportunity to meet peers. The idea of online social as well as professional networking is not by any means new. It's enough to look at existing virtual networking giants such as Facebook (www.facebook.com) and LinkedIn (www.linkedin.com). The available networking mechanisms, however, have a very broad scope. We are planning to provide a networking platform concentrated on CI and ML field. Such a platform will be a medium of communication for CIML researchers, students, teachers and anybody else interested in CIML. It will also have other functions common to popular networking sites such as profile pages, job searching, etc.

2.6 Integration of all components

Each of the components presented above has a very significant role by itself.

However, the true strength of the CIML community will be a deep integration of all of these components. Did you ever feel like discussing your initial results on the dataset or software that you just downloaded? Or after finding a piece of software, you are interested in finding related tutorials literature references, or even a link to the developers profile so you can ask him/her questions? Were you ever impressed by a piece of work and you wanted to offer the author a position in your lab? We believe that our unifying approach will make all of this much easier.

3. CIML Community Portal

The main platform in which we are implementing all the ideas presented above is CIML Community portal. It can be found at http://www.cimlcommunity.org. This section briefly presents our initial efforts in building the portal.

3.1 Current development stage

Figure 1 presents the main page of the portal. Following well-known rules of developing user-friendly websites, our goal is to create a simple and clear graphical interface. The real challenge of the portal is to link all relevant information together, without adding unnecessary complexity to its form. Currently the portal is in its development stage, providing the site for submitting as well as downloading software. It also includes materials presented at the Workshop on *Building Machine Learning and Computational Intelligence Virtual Organizations* held at George Mason University on October 24, 2008.

To submit a program, the user has to register to our portal as a Developer and fill out a simple form before uploading files. Currently we do not restrict the language used for software development. Also we accept open source as well as executables only. As soon as a program is submitted, it is subject to a peer review, similar to scientific journals. When accepted, the program is put on the portal and is immediately accessible to all Users. In addition, all posted items are to be citable resources. For proper referencing, the postings will include titles, name of the authors (or developers) and publishing dates. This will give due credit to all developers-contributors. With the growth of the portal, the submitted resources will be appropriately categorized and linked together to offer a meaningful content.

An example of software available on our portal is Adaptive Linear Hyperplane (ALH) for Classification, developed by Tao Yang and Vojislav Kecman at the University of Auckland. The program implements a new classification method proposed very recently in[11] by the program authors. It is implemented in MATLAB. The program allows for applying the method to various classifications problems and should allow other researchers to be able to compare its performance to the performance of their methods, as well as extend Yang's and Kecman's method.

3.2 Who is behind it?

The virtual community development team consists of three professors from the University of Louisville and George Mason University, as well as three students from the University of Louisville.

Our community also has 25 members, i.e. individuals that serve an advisory function in building the community as well review software etc. They are mostly well renowned researchers in computational intelligence and machine learning from the United States (13 members) and other parts of the world (12 members). The full list of members is on the portal web site.

4. Conclusions

The future of science is in collaboration between groups and scientists distributed all over the world. Our vision is to create a computational intelligence and machine learning virtual community that brings together scientists, educators, students, and all others interested in this discipline. In this paper we described benefits and opportunities as well as difficulties that need to be overcome when building a virtual community in computational intelligence and machine learning. We discussed general aspects of such a community and presented our initial and ongoing work on their implementation. However, a community will not exist without members. Therefore, we would like to encourage readers

FIGURE 1 The main page of CIML Virtual Community portal.

(continued on page 54)

Students choose an aspect or application of interest and investigate it using one or multiple paradigms. They must make a literature review on their topic. The project deliverables are a written report including the code and an oral presentation to the class. The assessment is based on both components and focuses on technical content and comprehensiveness. For the oral component, students must adhere to their stated time or be penalized. They are also assessed on how they respond to questions during the presentation (though this is omitted for the outreach students as they submit their presentation as a video). For the written report, the grade depends on the technical scope, the suitability of the method(s) to the problem(s), the appropriateness and innovation of the method(s) as coded, and the basis of assessment and comparison of the method(s) on the problem(s). In the presentation of results, the succinctness while still including all important information is important, and quite difficult for students to achieve. Students tend to want to show all runs of all trials, usually in voluminous tables. This does not lead to an effective presentation. So, one of their learning objectives is to decide how to best present results and which results to present. Over the years, some of these classroom projects have led to publishable work. (In fact, my most cited paper began as a class project in this course about 15 years ago [6]).

Course Documentation

This consists of syllabus, PowerPoint slides, homework assignments, journal papers to read and discuss, project assignments and the examination. During the course, these are posted to the course web site (except for the examination) for students to access. These are available by request from the author (smithae@auburn.edu). I have used mainly Reeves' text, *Modern Heuristic Techniques for Combinatorial Problems* [7], for this course. However, this excellent book is now out of print. The most recent semester of this course used *Metaheuristics for Hard Optimization* by Dreo, et al. [8] as the text. Both texts have been supplemented quite a bit. For example, the Reeves text does not include evolutionary strategies, ant colony methods nor particle swarm optimization.

In summary, this course is comprehensive and addresses the aspects of adaptive optimization of most use to technical students solving complex system design problems. The level of coding needed for this course and the amount of time required to code each paradigm and carry out the computational work are barriers to students enrolling in the course. However, there is no other way to deliver a meaningful course without this stringent reliance on coding and computation so I have chosen to live with relatively low student numbers in favor of the quality of the experience for each student.

References

[1] X. Yao, "A research-led and industry-oriented MSc program in natural computation," *IEEE Comput. Intell. Mag.*, vol. 1, no. 1, pp. 39–40, Feb. 2006.

[2] U. Kaymak, "An MSc program in computational economics with a focus on computational intelligence," *IEEE Comput. Intell. Mag.*, vol. 1, no. 2, p. 41, May 2006.

[3] N. Metropolis, A. W. Rosenbluth, M. N. R. A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, June 1953.

[4] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, May 1983.

[5] E. Bonabeau, M. Dorigo, and G. Theraulaz, "Inspiration for optimization from social insect behavior," *Nature*, vol. 406, pp. 39–42, July 2000.

[6] D. W. Coit and A. E. Smith, "Reliability optimization of series-parallel systems using a genetic algorithm," *IEEE Trans. Rel.*, vol. 45, no. 2, pp. 254–260, June 1996.

[7] C. Reeves, *Modern Heuristic Techniques for Combinatorial Problems*. New York: Wiley, 1993.

[8] J. Dreo, A. Petrowski, P. Siarry, and E. Taillard, *Metaheuristics for Hard Optimization*. Berlin: Springer, 2006.

Research Frontier *(continued from page 46)*

to visit our portal (http://www.cimlcommunity.org) and become both active Users and Developers. Furthermore, if the topic meets your interest and you would like to participate in creating the CIML Community, you are invited to do so. Just contact us at cimlcommunity@louisville.edu

Acknowledgments

This work is supported in part by the National Science Foundation Grant No. CBET 0742487. The findings and opinions expressed here are those of the authors, and do not necessarily reflect those of the sponsoring organization. The authors would like to thank Jordan Malof for his help in preparation of this manuscript.

References

[1] J. Preece, C. Abras, and D. Maloney-Krichmar, "Designing and evaluating online communities: Research speaks to emerging practice," *Int. J. Web Based Communities,* vol. 1, no.1, pp. 2–18, 2004.

[2] G. DeSanctis and P. Monge, "Communication processes for virtual organizations," *Org. Sci.*, vol. 10, no. 6, pp. 693–703, 1999.

[3] J. Cummings, T. Finholt, I. Foster, C. Kesselman, and K. A. Lawrence. (2008). "Beyond being there: A blueprint for advancing the design, development, and evaluation of virtual organizations," *The final report from the workshops on Building Effective Virtual Organizations* [Online]. Available: http://www.ci.uchicago.edu/events/VirtOrg2008/VO_report.pdf

[4] I. Foster, C. Kesselman, and S. Tuecke, "The anatomy of the grid: Enabling scalable virtual organizations," *Int. J. High Perform. Comput. Applicat.,* vol. 15, no. 3, pp. 200–222, 2001.

[5] C. Catlett, J. Boisseau, J. Cobb, T. Cockerill, M. Levine, M. Sheddon, C. Song, R. Stevens, and C. Stewart. (2007, Jan.). NSF Extensible Terascale Facility TeraGrid, *TeraGrid Annual Report and Program Plan* [Online]. Available: http://www.teragrid.org/about/docs/TG-Annual-2007-Pub.pdf

[6] S. Dong, J. Insley, N. T. Karonis, M. E. Papka, J. Binns, and G. Karniadakis, "Simulating and visualizing the human arterial system on the TeraGrid," *Future Gener. Comput. Syst.*, vol. 22, no. 8, 2006.

[7] P. Bonetto, G. Oliva, A. R. Formiconi, "MedIGrid: A medical imaging environment based on a grid computing infrastructure," in *Proc. 25th Annu. Int. Conf. Eng. Med. Biol. Soc.*, pp. 1338–1341, 17–21 Sept. 2003.

[8] I. H. Witten, and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. San Fransisco, CA: Morgan Kaufmann, 2005.

[9] B. Sendhoff, M. Roberts, and X. Yao, "Evolutionary computation benchmarking repository," *IEEE Comput. Intell. Mag.*, pp. 50–52, Nov. 2006.

[10] J. M. Zurada, J. Wojtusiak, F. Chowdhury, J. E. Gentle, C. J. Jeannot, and M. A. Mazurowski, "Computational intelligence virtual community: Framework and implementation issues," in *Proc. Int. Joint Conf. Neural Networks (IJCNN 2008)*, 1–6, June. 2008, Hong Kong, pp. 3152–3156.

[11] T. Yang and V. Kecman, "Adaptive local hyperplane classification," *Neurocomputing*, pp. 3001–3004, Aug. 2008.