Toward Multidisciplinary Collaboration in the CIML Virtual Community

Jacek M. Zurada, Janusz Wojtusiak, Maciej A. Mazurowski,
Devendra Mehta, Khalid Moidu, Steve Margolis

***Abstract*—The importance of virtual scientific communities is constantly growing, as they provide opportunity for collaboration between members located around the world. The Computational Intelligence and Machine Learning (CIML) Virtual Community aims at providing resources to researchers, students, and general public interested in the area. This paper describes how the CIML Virtual Community can support collaboration between CIML and researchers in the healthcare profession. It also describes how it can be generalized to support other disciplines interested in applying CIML methods.**

***Index Terms*—Collaboration, Computational Intelligence, Machine Learning, Virtual Community**

I. Introduction

Virtual communities play an important role in modern science. They allow community members to quickly exchange information across the globe. A *virtual scientific community* is a group of people, often researchers and students, who share multiple resources related to the scientific field, and whose main medium of communication is the Internet. Communication between members often requires creating web portals used to host the resources. While sharing resources is crucial, it is not sufficient for the existence of a virtual community. For example, many web portals that provide community resources, e.g. lists of publications, are not virtual communities. One example of such a portal is DBLP [1], which is one of the most popular lists of publications in Computer Science. Because there is no actual group of members that collaborate through DBLP, it cannot be considered a virtual community. Probably the best organized contemporary virtual communities are those oriented towards specific topics in medicine, bioinformatics, and related areas. This is due to the requirement (in those fields) that all material must be submitted to a well established repository in order to be considered for publication. An example of such a community is the Biomedical Informatics Research Network (BIRN), which provides access to data, tools, and collaborative infrastructure [3].

Wikipedia [2] is also a great source of information in many domains. It follows the ideas of Web 2.0, in which users create content for the web. This content is, however, sometimes not well organized and incomplete. Although most articles are intended for the general public, some of them are very technical and can be understood only by experts. The advantage of Wikipedia is that each article can be modified by multiple authors and therefore reflects their diverse and expansive body of knowledge. The downside of using Wikipedia is that the information is often unverified and expresses personal opinions of the authors which may not necessarily reflect those of the scientific community. In contrast, papers published in scientific journals most often require strict review process which significantly increases the reliability of the published material.

Computational Intelligence and Machine Learning (CIML) is a rapidly growing discipline. Despite the relatively mature state of virtual communities, the idea of building a global CIML community is unfortunately very new. As a result, the current status of virtual cooperation within CIML is worse than in many other domains. Although there have been several attempts to create collaboration websites, data and software repositories, and actual virtual communities, these efforts haven't been coordinated, and respond only to a few specific aspects of CIML community needs as a whole. Among the most noticeable efforts in building CIML virtual communities are the PASCAL and PASCAL2 networks [2] which are European initiatives that support collaboration and research in cognitive systems. Particular areas of interest of the network within CIML include machine learning, pattern analysis, machine vision, and natural language processing. Despite the fact that the networks' website is rich in content (e.g. publications, video lectures, competitions) many items are available only to its members.

Most CIML resources are distributed over countless websites maintained by single researchers, groups, laboratories, and departments. These websites are usually focused on specific topics of interest and leave little room for

This work was supported in part by the National Science Foundation Grant CBET 0742487. The findings and opinions expressed here are those of the authors, and do not necessarily reflect those of the sponsoring organization.

J.M. Zurada is with the University of Louisville, Louisville, KY 40292 USA (phone: 1 502-852-6314; 1 502-852-3940; e-mail: jacek.zurada@louisville.edu).

J. Wojtusiak is with the George Mason University, Fairfax, VA 22030, USA (phone: 1 703-993-4148; e-mail: jwojtusi@gmu.edu)

M.A. Mazurowski is with the University of Louisville, Louisville, KY 40292 USA (phone: 1 502-852-3165; fax: 1 502-852-3140; e-mail: maciej.mazurowski@louisville.edu).

K. Moidu is with the George Mason University, Fairfax, VA 22030, USA and Orlando Health, Orlando, FL 32806 (phone: 1 703-993-9734; e-mail: kmoidu@gmu.edu)

D. Mehta is with Orlando Health, Orlando, FL 32806 (e-mail: Devendra.Mehta@OrlandoHealth.com)

S.S. Margolis is with Orlando Health, Orlando, FL 32806 (e-mail: smargoli@OrlandoHealth.com)

any comprehensive or broad view of the field. In addition, most of these sites also lack any objective content evaluation. Probably the most well known websites in CIML is one which hosts implementation of some standard machine learning algorithms in Java™ within the Weka system [5]. The software is available from the University of Waikato website [6]. Another example of a very popular site is UCI machine learning repository with collection of benchmark data [7].

The above typically concentrate on a single aspect of collaboration, while a more global look at a CIML virtual community would provide not only access to its particular components or functions (such as data sharing or networking). It would also create an interconnection between the components making such a community more integrated and better informed. In this paper, we present our initial efforts towards implementing CIML virtual community, and present how it may lead to multidisciplinary collaboration across disciplines [8].

II. Computational Intelligence and Machine Learning Virtual Community

A. Role of the CIML Virtual Community

As the fields of computational intelligence and machine learning mature, there is a growing need to provide researchers with the ability to exchange information, share resources, discuss problems and new directions, and learn about others' work. In the past, scientific journals were the most important medium of communication between researchers. In the rapidly changing and very dynamic field of CIML, this form of communication is simply too slow for everyday exchange of information. Very quick review processes still take months. In many cases, particularly for high quality journals, it may take two or three years between the original submission and the actual publication. Professional conferences provide the opportunity to meet other researchers as well as present and discuss results. With shorter

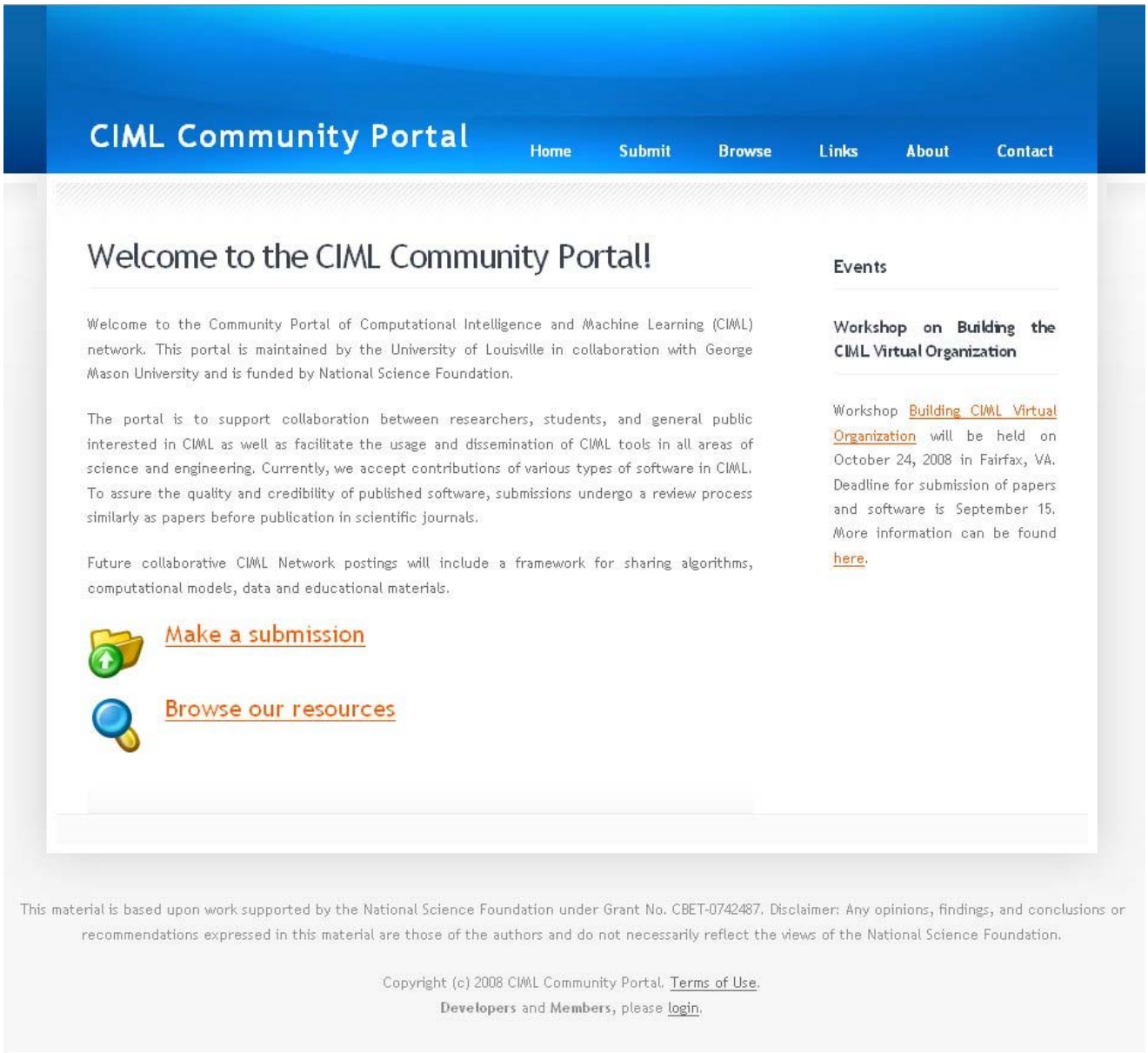

Figure 1: The main page of CIML virtual community portal.

review processes, often in the order of a few months, these conferences allow more rapid communication and discussion of research results. Despite these benefits however, high travel costs often prevent potential attendees, in many cases students and distant researchers, from attending.

The aforementioned limitations, along with others, of traditional scientific communication inspired us to create a CIML virtual community. The goal of the community is to create a place where scientists, students, and the general public can work together despite any of their geographic limitations. The next section briefly describes our initial efforts to create such a community and presents its current status.

B. Current Status

The CIML virtual community board membership currently consists of 25 people, which are well established researchers in the area. Thirteen members are form the United States and twelve are from other countries. These members help in building the community and its various components.

Currently the main initiative is building The Computational Intelligence and Machine Learning community portal. The portal will serve as a medium to exchange data and software, as a professional networking platform, and as a source for help in obtaining educational materials. A screenshot of the main page of the portal is presented in Figure 1. The portal's core development team consists of two professors from the University of Louisville, one professor from George Mason University, and three students from the University of Louisville. Despite its youth, the community is already accepting submissions of CIML software. The software goes through a review process similar to that used by scientific journals, and upon its acceptance is published on the portal.

III. Collaboration with The Medical Community

A. Goals

Healthcare is an area with diverse problems, types of datasets, and study objectives. Researchers in the medical and general healthcare domains frequently use popular statistical methods, but are not familiar with the wide range of methods and tools available in CIML. Moreover, current physicians in training are asked to do research. Also the challenge of search for solutions drives the senior physicians to conduct research. The power of a collaborative would empower them to compare the data sets and results they have with others researching similar issues around the core problem.

National collaborations in Oncology, for example, have shown the importance of networks. These collaborations are, however, only possible in a few heavily funded fields. Partly owing to the success of such collaborations, areas like clinical research increasingly depend on multicentered studies. Networks that allow collaborations in similar or overlapping areas would be particularly advantageous and allow large collaborations to develop, especially in areas not traditionally well funded. Access to the CIML virtual community would facilitate this process tremendously by connecting individual physicians, assigned representatives, and geographically distributed experts. Even down to individual practices, an environment of ongoing case review and outcome analysis leads to an environment of data driven changes, which is advocated by multiple organizations including the Institute of Medicine. These measures are often taken on a small scale, yet improvements have lead to their having substantial impacts. Application of advanced analytical tool may still not necessarily occur, however, as resources or research staff may be lacking knowledge or resources. Useful presentations and publications of this type may not be in the mainstream journals of the discipline. Collaboration within the CIML virtual community could lead to larger scale review of similar cases or scenarios, a more robust meta-analysis of data, which in turn could lead to more defined strategies to improve outcomes. Consequences of such a large, open, and more diverse community would include availability of immediate feedback from peers, as well as immediate dissemination of successful strategies. Therefore, research results obtained by using CIML in healthcare have the potential for a very high impact.

Concluding from the discussion above, healthcare is a domain in which CIML tools and virtual collaboration could yield significant results.

B. Examples of Possible Applications

The first example concerns managing nosocomial infections, an important step in reducing overall morbidity and mortality rates. By the careful logging of critical variables, measures undertaken, and incidence of such infections, comparisons in specific patient populations, institutions, or approaches would be possible on an ongoing manner and in real time. Moreover, measures associated with lower rates can be readily identified. Sharing of such data would allow powerful analysis on large sample sizes, and identify potential risk factors that would otherwise not be recognized. While identification of outstanding institutions and approaches would occur readily, subtle gains as well as major breakthroughs from a myriad of approaches could also be more quickly identified and adopted by others. Therefore, application of CIML tools available within the virtual community would allow previously impossible analysis of nosocomial infections data.

Another example concerns reflux disease and asthma. These two diseases commonly coexist. Recently, the ability to identify acidic ($pH<4$), weakly acidic ($pH<4\leq7$) or alkali ($pH>7$) reflux and level up to the upper esophagus has been possible using combined impedance and pH probe studies. Furthermore, the ability to detect microaspiration has improved with the recent advent of airway pepsin as a biomarker of gastric aspiration. The potential role of pepsin as a cause of irritation, inflammation needs to be evaluated along with other gastric fluid constituents or properties including acidity, hypotonicity, and even bile or microbes. Using cough as a defined event in asthma, we studied 117 children with asthma to assess pH of refluxate immediately prior to a cough.

In a subset who had bronchoscopy, we assessed the prevalence of pepsin positivity in the airway. There were 27 attributes in the core study. Some attributes related to severity of asthma, based on need of medications to control asthma, and measures such as spirometry. This is a composite score because with young children, spirometry is not available, and their classification may be less robust. Additional attributes emanate from impedance pH probe studies. These are studies carried out over 18 to 24 hours, with or without acid suppression. We look at characteristics of reflux, as well as correlation with any symptoms. These results are collected by the physician after manual review or electronically from the software. Specific elements are entered into a database. Several attributes help relate characteristics of reflux and subsequent symptoms. Normative scales are used to assess categorical interpretation of the studies as normal or abnormal by the physician. Other attributes collected include symptoms that suggest gastrointestinal reflux, evaluation results of any tests for reflux, and pulmonary symptoms over time. These are collected at bedside and entered in an electronic medical record. The relevant information is then parsed and transferred to a database. In a subset who had undergone bronchoscopy, description of the airway in terms of inflammation, narrowing or other lesions, as well as data from airway washings for pepsin activity were collected. This data comes from studies performed by pulmonologists at a separate time point to impedance study, and are collected via chart review (these are performed at separate institutions with varied electronic record systems, hence the need for a manual review and entry). For part of the data collection, entry into paper forms was performed, which can later be scanned into an electronic format.

By using a CIML approach, predictors and determinants can be derived from additional historical information. In prospective phases where interventions are planned, the multiple factors affecting asthma would be possible. These include psychosocial, seasonality, allergy, and a clinical course over a lengthy timeline that would allow small changes to be detected. Such an approach could also elucidate the potential role of reflux and microaspiration on wheezing infants and toddlers, interstitial lung disease, and other chronic lung diseases. Likewise, potentially significant results can be obtained in patients with poorly protected airways, including in national and pediatric intensive care unit populations. Finally, the role of pepsin in other etraesophageal manifestations of reflux, such as laryngitis, otitis media, and perhaps sinuses could be elucidated. As a result, predictive models could be created to offer point of care decision support.

C. Methodology

Healthcare researchers interested in applying CIML tools to solve problems in their field of study can utilize tools, methods, and expertise available at the community's portal. In general, we can assume that they have some data to be analyzed. Such data can come in many different forms: structured (from existing databases), text, images, time series (e.g. EEG), and others. There is often a background knowledge associated with the particular domain in the form of prexisting models, rules, ontologies, dictionaries, hierarchies, etc.

Once the data and possible background knowledge is available, the user needs to define the problem of interest. This is one of the most difficult parts of the process as it requires mapping of sometimes very vague problem description into the CIML methodology. For example, the statement "I want to find out what are possible reasons for chest pain among my patients" is too imprecise to be directly addressed by CIML tools. It needs to be translated into the right problem such as classification, clustering, or optimization. This process should be supported by a series of questions that correspond to entries illustrated in the Figure 2**.**

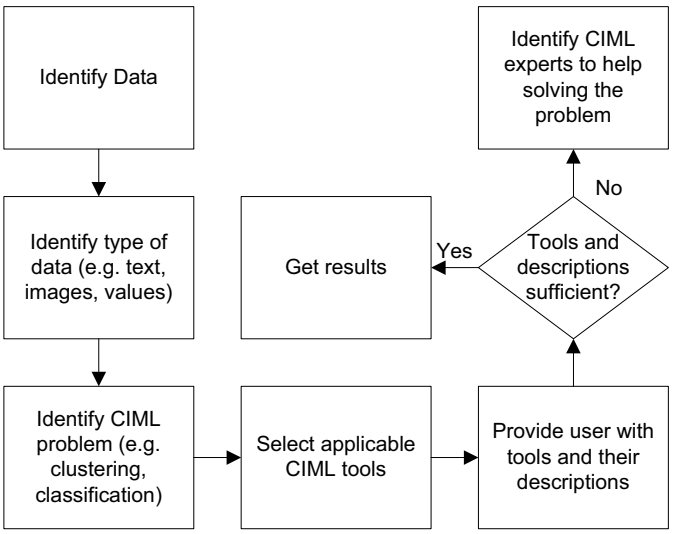

Figure 2: Steps in selecting CIML tools.

Interaction with the CIML community portal should lead users to:

- Selection of *relevant* CIML tools
- Access to *relevant* tutorials and articles
- Contact information for CIML community members whose area of expertise is *relevant* to the considered problem.

The key word in the above is "relevant" because the portal should guide users through its resources. The healthcare researchers may then access and apply the selected tools to obtain actual results. The complete process is illustrated in Figure 3. Because a significant part of the data in medical records is stored in the form of text (e.g. diagnoses information), we emphasize it in the below diagram. Using text recognition tools it can be transformed into structured data that's viable for further analysis.

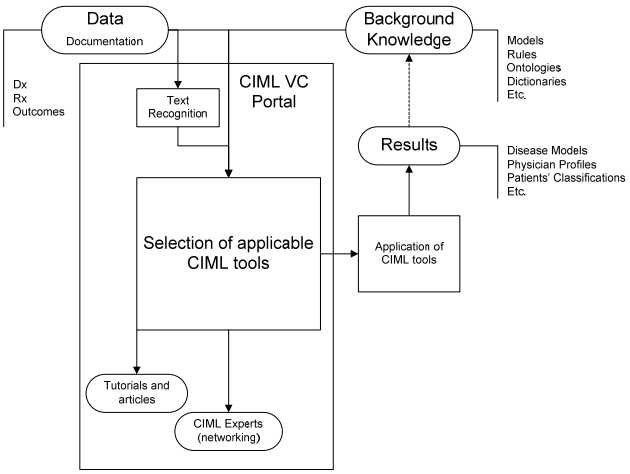

Figure 3: Diagram of a general methodology for choosing applicable CIML tools in medical domain.

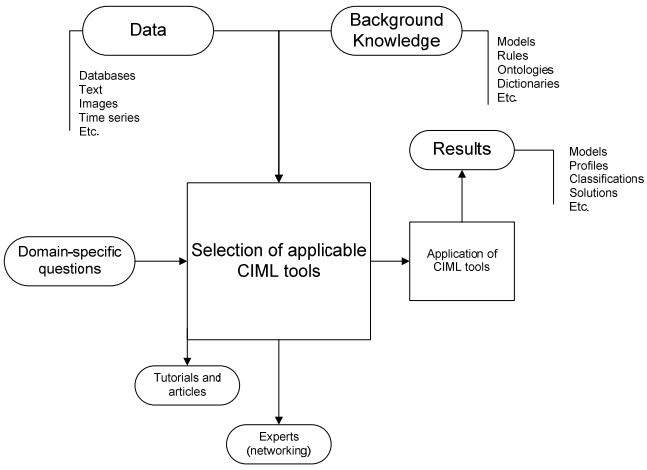

Figure 4: Diagram of a general methodology for choosing applicable CIML tools across disciplines.

IV. Collaboration with Other Disciplines

While the medical domain provides an important area of application of CIML tools, there is also great potential for the CIML community to impact other disciplines. The fact that scientists from various disciplines may be interested in applying methods developed within CIML does not require much justification. It is sufficient to look at the content of conferences and journals in both CIML and other disciplines to discover a wide range of current and potential applications. The methodology for collaboration between the CIML community and health care researchers presented in the previous section can be generalized to other disciplines as depicted in Figure 4. The key part is generalization of the most important modules, namely selection of applicable CIML tools and experts. Creation of such a module to work across disciplines will require creating an adaptive methodology that will be able to automatically incorporate new community members and software tools submitted to the community.

The methodology for enabling collaboration between CIML and any other discipline is the same. The challenge is to adapt terminology used by the portal to what is easily understood by users at different levels of CIML knowledge. Researchers in different fields, or even in the same field, tend to use different terms when talking about the same things.

A potential solution to the above problem is to create ontology of terms used in computational intelligence and machine learning, and then map it onto terms used in different disciplines. Thus, experts from different disciplines would be asked questions using their own domain-specific language.

V. Conclusion

Benefits from multidisciplinary collaboration within computational intelligence and machine learning are numerous. Researchers working with multiple disciplines clearly benefit from the access to state-of-the-art CIML tools, their descriptions, articles, and researchers. On the other hand CIML researchers benefit by having the possibility to drive research by real world problems. It is our intention to initiate and support multidisciplinary collaboration whose central part is development and use of CIML tools and methodologies.

Acknowledgement

The authors thank Jordan Malof for his valuable comments that helped to improve this paper.

References

[1] M. Ley, DBLP website, http://www.informatik.uni-trier.de/~ley/db/
[2] Wikipedia website: www.wikipedia.org
[3] The Biomedical Informatics Research Network (BIRN) http://www.nbirn.net
[4] The PASCAL network website: www.pascal-network.org
[5] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Series, 2005.
[6] www.cs.waikato.ac.nz/~ml/
[7] The Machine Learning Repository, University of California, Irvine website: http://archive.ics.uci.edu/ml/
[8] 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*, International Joint Conference on Neural Networks (IJCNN 2008), June 1-6, 2008, Hong Kong, pp. 3152-3156.