Intelligent Information Systems 9999
ISBN 666-666-666, pages 1–12

Toward VO-based Collaboration between Computational Intelligence - Machine Learning and Healthcare Communities

Janusz Wojtusiak[1], Jacek M. Zurada[2], Jordan M. Malof[2], Devendra Mehta[3], and Khalid Moidu[1,3]

[1] George Mason University, USA
[2] University of Louisville, USA
[3] Orlando Health, USA

Abstract

Computational Intelligence - Machine Learning (CIML) is a mature discipline with a wide range of potential applications. Many state-of-the-art methods developed in CIML, however, remain unused outside the relatively small groups of researchers in which they originate. This can be attributed to lack of communication between disciplines, difficulty of use, availability, and limited understanding of CIML among members of other communities. Healthcare is a discipline that may particularly benefit from using CIML methods because of its complexity and significance. An effective methodology for collaboration between CIML and healthcare communities would create a beneficial opportunity for both sides. One possible methodology, based on automated advising, is described here in the context of the Computational Intelligence and Machine Learning Virtual Community. The community's main objectives are to unite efforts in the field, provide resources for researchers, educators, and the general public as well as reach out to other disciplines.

Keywords: Collaboration, Computational Intelligence, Healthcare, Machine Learning, Virtual Organization

1 Introduction

The field of computational intelligence and machine learning has reached a mature state. It is an extensive field covering a wide range of methods with applications to many types of problems. It is, thus, surprising that many of its methods are either used only by their authors, or published and "shelved" without much wider interest. The situation is even worse with computer programs and software developed in the course of the research. These materials are rarely available for public use, or they are sometimes on websites maintained by authors where in most cases they are only known and accessible by small groups of people. This makes it difficult for members of other disciplines to use these materials in their studies. Many great methods and tools developed in CIML are simply wasted.

Several websites related to CIML exist. One such place that many of us often visit is Wikipedia (www.wikipedia.org). It hosts many articles related to CIML, and it is often a good starting point when trying to find information, even for experienced CIML researchers. There are, however, important weaknesses of the open approach in which users freely edit content. Its content quality control is limited (we can only assume that when one user writes something incorrect, others will fix it), levels of difficulty of articles are mixed (some basic information for beginners is mixed up with details for experts), and it is hard to reference since it changes over time. An initiative whose goal is to address some of the weaknesses is Scholarpedia (www.scholarpedia.org), in which articles are peer-reviewed to ensure their quality. Similar to Wikipedia, articles may be edited by the "community," but to ensure quality all changes need to be approved by editors or peer-reviewers.

Probably the most common place for hosting computer programs is Sourceforge (www.sourceforge.net) which, unfortunately, also lacks quality control over its content and is mostly targeted towards programmers and experts in specific domains. Despite the popularity of these sites, they are insufficient in providing CIML resources and knowledge to outside communities who might benefit from applying these methods.

This paper tries to answer the question of how members of other communities can easily navigate through innumerable CIML methods and select the most appropriate software tools. While this discussion concentrates on the healthcare community, because of its complexity and the variety of problems involved, it can be easily generalized to other disciplines. A description of the proposed methodology of collaboration is followed by a classification of CIML methods from the perspective of their use by other disciplines. Finally, we present current efforts in building the computational intelligence and machine learning virtual community (CIMLVC), which will host implementation of the methodology.

2 Collaboration with Healthcare Community

2.1 Goals

Healthcare is an area with diverse problems, types of datasets, and study objectives (Cios and Moore, 2002). These areas of interest range from predicting outcomes of treatments and classification of diseases to optimization of hospital workflows and financial evaluation of physicians. 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. This is also the case for physicians in training who are asked to do research. As a result of challenges faced, senior physicians conduct research as well. A collaboration with the CIML community would empower them to use state-of-the-art methods and compare the data sets and results they have with others researching similar issues around the core problem (Zurada et al., 2008b).

National collaborations in oncology, for example, have shown the importance of collaborative 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 larger collaborations to grow, especially in traditionally underfunded areas. Access to the CIML community tools and methods 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 tools 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 be absent in the mainstream journals of the discipline. Close collaboration with the CIML community could lead to larger scale review of similar cases or scenarios and 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.

An important issue that concerns application of any methods in heathcare, particularly when directly working with patients' data, is reliability of applied methods. The purpose of collaboration between CIML and healthcare communities is mainly research oriented, specifically in most cases the goal is to discover knowledge that needs to be later verified (e.g. through clinical trials). This process along with all privacy and security issues are strictly monitored by Institutional Review Boards at institutions where the research is conducted.

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

2.2 Examples of Possible Applications

The first example concerns managing nosocomial infections, an important step in reducing overall hospital morbidity and mortality rates. By the careful logging of critical variables, measures undertaken, and incidence of such infections, comparisons of specific patient populations, or approaches at other institutions, could be performed on a regular basis and in real time. Consequently, factors associated with lower mortality and morbidity rates could 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 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 ($4 < pH \leq 7$) 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 of children who had bronchoscopy, we assessed the prevalence of pepsin positivity in the airway.

There were 27 attributes in the core study. Some attributes were related to the severity of the asthma, based on the need of medications to control it, as well as 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 of patients 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 points, 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.

2.3 Methodology

Healthcare researchers interested in applying CIML to solve problems in their field of study should be able to utilize tools, methods, and expertise in the area. Currently no such possibility exists without direct access to CIML experts. We suggest a solution in which this expert role is taken by an automated system

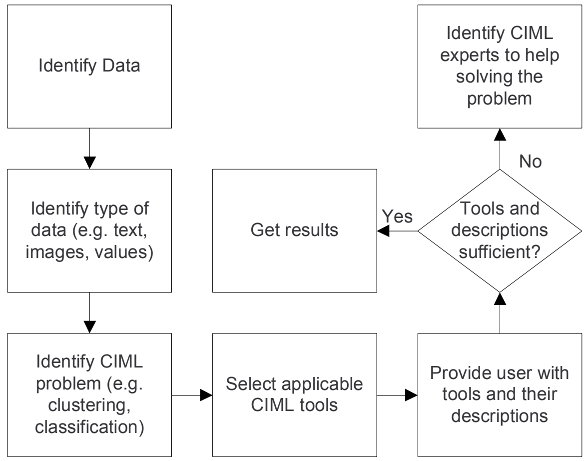

FIGURE 1: Steps in selecting CIML tools.

within the CIML virtual community portal, whose initial version is described in next section. While the presented methodology is not yet fully implemented, it is emerging as a part of the virtual community. Because of its complexity the project (particularly creating ontologies and collections of resources) may take several years and requires broad participation of the CIML community.

For discussion of the collaboration methodology, let us assume that a potential user (e.g. healthcare researcher) has some data to be analyzed. The methodology can easily be generalized to other types of problems (e.g. optimization in which instead of data an objective evaluation model is given). 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 preexisting models, rules, ontologies, dictionaries, hierarchies, correlations, 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 descriptions into the CIML methodology. For example, the statement "I want to find 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 1.

Depending on their level of experience and the considered problem, users should be given access to different types of available resources.

- **CIML software and tools.** A wide range of computer programs developed in CIML is available. These programs, however, are often incompatible, platform-dependent, or hard to use for non-technical people. Selected programs not only should be relevant to the problem being addressed, but also have standardized input and output, and be easy to use. Because programs have different capabilities, can analyze different types of data, and are controlled using different parameters, it is suggested that XML may serve as a common language for describing data (Zurada et al., 2008a).

- **Tutorials and articles.** In addition to user's guides and manuals that describe technical aspects of using software and tools, users from other communities should have access to articles and tutorials that help them understand the advantages and disadvantages of specific methods. The documents should be linked to specific methods relevant to them. Thus, one can view documents relevant to specific software (e.g. when accessing a rule learning program a selection of documents describing concept learning, rule learning, and the specific methods should be available), or when accessing specific articles a list of software should be available (e.g. the article describing neural networks should be linked to available neural programs). It is also important that each document is categorized based on its difficulty. Difficult technical descriptions for experts should not be mixed with simple tutorials and descriptions for non-experts.
- **CIML experts.** It is not always possible to obtain satisfactory results without consultation with CIML experts. In such a case it is important to be able to identify such experts and provide contact information for them. This may lead to possible collaboration that would not be possible otherwise, and which can be mutually beneficial - CIML community gains access to real world problems that can spark novel development and research directions. Obviously members of other communities benefit from expertise for their problems.

The healthcare researchers may then access and apply selected tools to obtain actual results. The complete process is illustrated in Figure 2. In the most common scenario, healthcare data is given in the form of diagnoses (Dx), medications (Rx), and outcomes. Because a significant part of the data in medical records is stored as text (e.g. diagnoses information), we emphasize it in the diagram below. Using text recognition tools it can be transformed into structured data that's viable for further analysis.

To achieve the automatic guidance through software, tools, articles, and datasets, each of these resources is described using a set of attributes. These attributes describe the type of resource, the applicable class of problems (see section on classification of CIML methods), level of needed experience, compatibility (e.g. specification of data formats compatible with a specific software), and other attributes specific for a given type of resource. The two most important classifications, namely methods and resources, are described in next section.

3 Classification of CIML Methods and Types of Resources

There are many possible ways of classifying methods in CIML. Methods can be classified based on representations, types of algorithms, types of input data, and so on (Atanassov and Hadjiski, 2008). The goal of the presented classification is to allow mapping of healthcare problems into CIML methods and tools. Thus, the classification presented in Figure 3 is based on types of problems to which these methods are applicable.

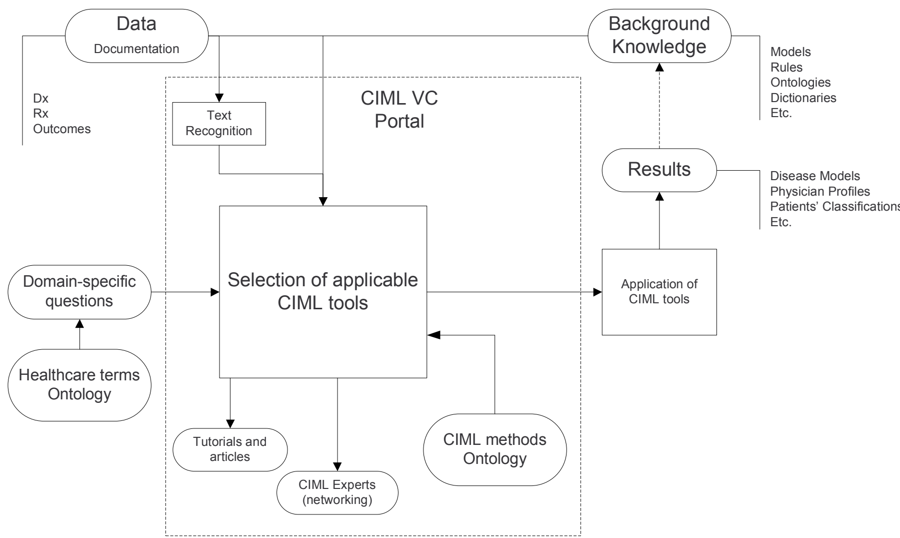

FIGURE 2: Diagram of a methodology for choosing applicable CIML tools in healthcare.

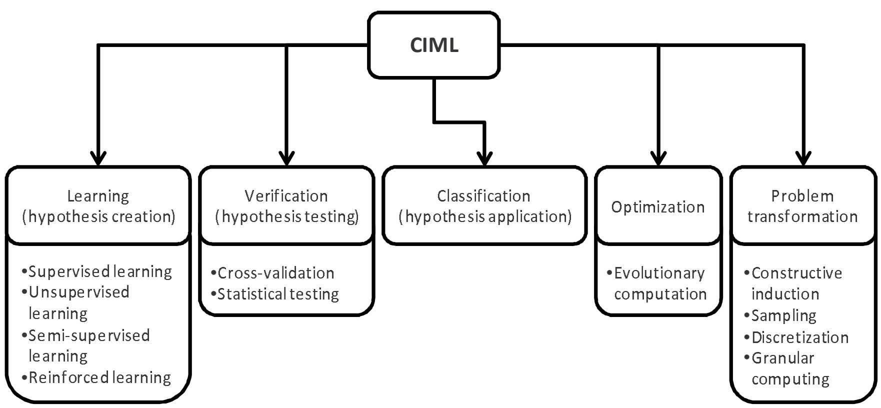

FIGURE 3: Classification of CIML methods.

It is important to note that much of the available computer software falls into more than one category in the presented classification. For example, most concept learning programs such as C4.5 (Quinlan, 1993), or LARS (Grzymala-Busse, 2005), are not only able to induce hypotheses, but also apply them to testing data. Other large systems, such as Weka (Witten and Frank, 2005) or VINLEN (Kaufman et al., 2007), are able to perform virtually any task mentioned below.

The general classification of CIML methods (Figure 3) is into five categories that correspond to major types of problems. Solving an actual healthcare problem involves possibly using tools from many of the above categories. For example a

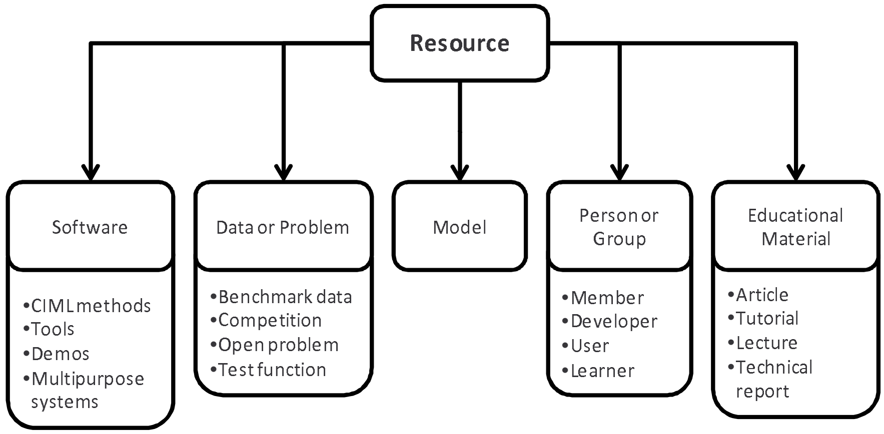

FIGURE 4: Classification of resources.

typical supervision learning problem involves data preprocessing (problem transformation), followed by actual inductive learning, then hypothesis verification, and finally application of learned knowledge. The above classification is not exhaustive, and is implemented in a dynamic form.

Resources are also classified into five major categories: software, data, problems, models, people, or educational materials (Figure 4). The classification is also not complete. Other categories, such as computing facilities, are possible, but are not yet considered because of their lower priority in the community.

Software is further categorized into CIML methods (learning, verification, classification, and optimization), tools (problem transformation, i.e. data preprocessing), demos (programs that illustrate use of specific methods), and multipurpose systems (programs that combine many of the above features). Data and problems include benchmark data and problems for testing new methods, problems posted as parts of competitions, and open problems for which the community seeks new solutions. Models resulting from the application of CIML tools also are valuable resources. They can be used for comparison with other models, or applied to real data. Educational materials include articles, tutorials, recorded lectures (possibly linked to existing sources such as those at www.videolectures.net), and technical reports (e.g. user's guides, technical descriptions of software, and other materials not publishable as regular papers). Finally, people or groups of people can be members that play the role of editors who review, accept, and manage resources. Other categories are developers who contribute resources, users who apply available resources to their problems, and learners (e.g. students or general public) whose sole purpose of accessing resources is to extend their knowledge in the field.

A specific resource can inherit properties from more than one category. For example an educational program for demonstrating rule learning belongs to both software (demos and CIML methods) and educational materials.

4 Computational Intelligence and Machine Learning Virtual Community

Virtual communities play an important role in modern science (Saltz et al., 2008). 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. Because there is no actual group of members that collaborate through DBLP (www.informatik.uni-trier.de/ ley/db/), 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 (www.nbirn.net).

Computational Intelligence and Machine Learning (CIML) is a rapidly growing discipline. Surprisingly, the idea of building a global CIML community is 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 (www.pascal-network.org) which are European initiatives that support collaboration and research in cognitive systems. Particular areas of interest of this 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. There is a number of initiatives or networks supported mainly by European Union with aims similar as the presented CIML portal. For example, Knowledge Discovery Network of Excellence (KDNet; http://www.kdnet.org), Knowledge Discovery in Ubiquitous Environments (KDUbiq; http://www.kdubiq.org). Another example of support for end-users of machine learning solutions is the MiningMart project (http://mmart.cs.uni-dortmund.de) that aims at building a library of data mining cases that can be reused for other projects.

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 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 website in CIML is one which hosts implementation of some standard machine learning algorithms in Java within the Weka system (Witten and Frank, 2005). The software is available from the University of Waikato website. Another example of a very popular site is UCI machine learning repository with collection of benchmark data available at website archive.ics.uci.edu/ml/.

The above example sites 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), but it would also create an interconnection between the components making such a community more integrated and better informed. Selected aspects of our initial efforts towards implementing the CIML virtual community are discussed below.

4.1 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 review periods, 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.

4.2 Current Status

The CIML virtual community board membership currently consists of 30 people, which are well established researchers in the area. Eighteen members are form the United States and twelve are from other countries. These members help in building the community and its various components. Our current initiative is the Computational Intelligence and Machine Learning community portal. The portal, when finished, 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 5. The portal's core development team consists of individuals from the University of Louisville and George Mason University. Resources (including software) submitted to the portal undergoes a review process similar to that used by scientific journals, and upon its acceptance is published on the portal (Zurada et al., 2009).

FIGURE 5: The main page of CIML virtual community portal.

5 Conclusion

Benefits from collaboration between computational intelligence and machine learning and other communities are numerous. One particular discipline that can achieve significant improvements is healthcare. Researchers working in this field clearly benefit from 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. Such collaboration requires, however, creation of an automated system that is able to guide its users through resources, and select the most appropriate methods of solving their problems. This paper briefly presented a methodology of automated collaboration between CIML and healthcare communities. The methodology is currently being implemented as a part of CIML Virtual Community, a joined effort by University of Louisville, George Mason University, and researchers from other institutions.

6 Acknowledgments

The authors thank anonymous reviewers for their suggestions on how to improve and extend the presented topic. This paper is a significantly extended version of work presented at the Workshop on Building CIML Virtual Organizations held at

George Mason University. 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.

References

K. Atanassov, M. Hadjiski 2008, Generalized nets as tools for modelling of intelligent systems, *Proceedings of the 4th International IEEE Conference on Intelligent Systems, IIS '08*, Varna, Bulgaria.

Krzysztof J. Cios, William G. Moore (2002), Uniqueness of medical data mining, *Artificial Intelligence in Medicine*, 26, 1–24.

Jerzy W. Grzymala-Busse (2005), LERS - A Data Mining System, *The Data Mining and Knowledge Discovery Handbook*, 1347-1351.

Kenneth A. Kaufman, Ryszard S. Michalski, Jaroslaw Pietrzykowski, Janusz Wojtusiak (2007), An Integrated Multi-task Inductive Database VINLEN: Initial Implementation and Early Results, *Lecture Notes in Computer Science*, 4747, 116–133.

CIMLVO (2008) *Proc. of the Workshop on Building Computational Intelligence and Machine Learning VOs*, George Mason University, Fairfax, VA, October 24, http://www.cimlcommunity.org/workshop/proceedings

John R. Quinlan (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers.

J. Saltz, T. Kurc, S. Hastings, S. Langella, S. Oster, D. Ervin, A. Sharma, T. Pan, M. Gurcan, J. Permar, R. Ferreira, P. Payne, U. Catalyurek, E. Caserta, G. Leone, M.C. Ostrowski, R. Madduri, I. Foster, S. Madhavan, K.H. Buetow, K. Shanbhag, E. Siegel (2008), e-Science, caGrid, and Translational Biomedical Research, *Computer*, 41, 11, 58–66.

Ian H. Witten, Eibe Frank (2005), *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Series.

Jacek M. Zurada, Janusz Wojtusiak, Fahmida Chowdhury, James E. Gentle, Cedric J. Jeannot, Maciej A. Mazurowski (2008a), Computational intelligence virtual community: framework and implementation Issues, *International Joint Conference on Neural Networks (IJCNN 2008)*, June 1-6, Hong Kong, 3152–3156.

Jacek M. Zurada, Janusz Wojtusiak, Maciej A. Mazurowski, Devendra Mehta, Khalid Moidu, Steve Margolis (2008b), Toward Multidisciplinary Collaboration in the CIML Virtual Community, Proceedings of the Workshop on Building Computational Intelligence and Machine Learning VOs, George Mason University, Fairfax, VA, October 24, 62–66. http://download.cimlcommunity.org/cimlweb/workshop/papers/10_Zurada.pdf

Jacek M. Zurada, Maciej A. Mazurowski, Artur Abdullin, Rammohan Ragade, Janusz Wojtusiak, James E. Gentle, Building Virtual Community in Computational Intelligence and Machine Learning, *Computational Intelligence Magazine*, 4, 1, 43–54.