SEARCHING AND REASONING WITH DISTRIBUTED RESOURCES IN COMPUTATIONAL INTELLIGENCE AND MACHINE

Janusz Wojtusiak[1], Jan Chorowski[2], Jaroslaw Pietrzykowski[1], Jacek M. Zurada[2]

[1] George Mason University 4400 University Drive, MSN 1J3, Fairfax, VA 22030, USA *(jwojt, jarek)@mli.gmu.edu*

[2] University of Louisville, Louisville, KY 40292, USA *(jkchor01, jmzura02)@louisville.edu*

Abstract

Virtual organizations (VO) are geographically distributed groups of people sharing common goals and willingness to collaborate. One of the important roles of virtual organizations is to facilitate sharing resources related to the area of collaboration. This paper presents an approach to handling resources in Computational Intelligence and Machine Learning distributed over a large number of sites. Resources are first discovered in the internet, evaluated, and then shared among users. Reasoning and adaptation methods can then be applied to best fit resources into users' needs without a long search process.

Key words: Virtual organization, resource, discovery, categorization, sharing

1 Introduction

The computational intelligence and machine learning (CIML) community, along with many other disciplines, is in a great need of realizing benefits from the latest technology trends and opportunities offered by internet resources. Spontaneous and organized virtual organizations (VO) emerge from within the community to better utilize available resources and foster joint efforts. Designing the CIML VO requires thinking about catering both to intra-communal, as well as extra-communal needs. Such needs are related to re-sources' warehousing, sharing, and use based on the particular user's needs and requirements.

One common task involves searching for available resources related to a specific problem. For example, researchers working on a new method may want to compare its efficiency with the state-of-the-art on a set of benchmark

tests. Unified repositories of datasets with well documented and verified solutions are of great help. Thanks to such repositories, researchers benefit from easy ways of comparing their methodologies against the state-of-the-art. Additionally, potential users of CIML tools are clearly provided with the advantages and disadvantages of various approaches for problems similar to their tasks at hand. Such repositories popularize CIML methods among professionals of other specialties. Real-life problems from other disciplines spark the creativity of CIML researchers and drive development of new methods, thus yielding benefits both within and outside the CIML community. Currently available repositories are, however, tailored towards CIML experts who are often proficient in programming, and thus are not appropriate for many potential users.

There are many examples of synergy between the CIML specialists and users of their technologies. For example, the Netflix Prize competition attracted attention of the CIML community from all over the world, including many young scientists researching in that area. New real world data set was made available for open competition. The Netflix company also benefited by having their problem solved.

Cooperation between the CIML scientific community – the technology providers – and members of other disciplines is highly valuable and often benefit all the parties involved. However, it is not an easy task. There is a substantial knowledge gap between both communities. The best available tools developed in CIML often have limited functionality and are proofs-of-concept rather than fully functional and user-friendly software. Further, there is a plethora of formats to store data and present results. A new user trying to use CIML methodologies may feel overwhelmed and forced to learn the intricacies of the tools instead of simply using them.

As the fields of computational intelligence and machine learning mature, there is a growing need to provide researchers with comprehensive infrastructure to exchange information, share resources, discuss problems and new directions, and learn about others' work. In the past, scientific journals and conferences were the major medium of communication between researchers. In the rapidly changing and very dynamic field of CIML, this form of communication becomes too slow for everyday exchange of information. Reviews of submitted journal papers may still take months. Professional conferences provide the opportunity to meet other researchers as well as present and discuss results. With somewhat shorter review periods which still take several months, these conferences allow faster 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. Even worse situations occur when members of other communities try to access state-of-theart CIML resources, contact expert CIML researchers for collaboration, search methods applicable to their problems, and so forth.

The main goal behind a CIML Virtual Organization is, thus, facilitation of intra- and interdisciplinary interactions and the creation of a point-of-entry for them. The VO should be flexible enough to guide potential users to the best available resources based on their needs, their knowledge of the CIML domain, computer literacy, and other factors. Such an approach would bridge the gap between providers and users of the computer-based intelligent tools.
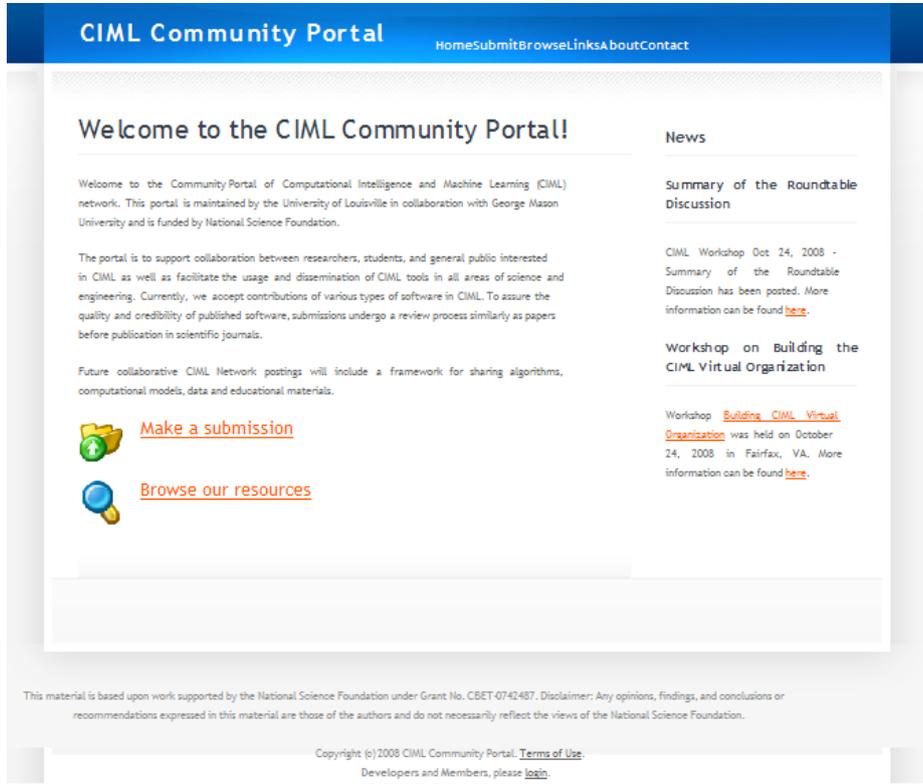
2   CIML Virtual Organization

Limitations of traditional scientific communication are driving forces behind the creation of the CIML virtual organization. The goal of the organization is to create a place where scientists, students, and the general public can collaborate despite geographic limitations. This section briefly describes our initial efforts to create such a community and presents its current status.

Currently, the CIML VO's board consists of 30 well established researchers in the area. The first initiative within the organization has been 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 1. The portal's development team consists of researchers from the University of Louisville and George Mason University. Resources (including software) submitted to the portal undergo a review process similar to that used by scientific journals, and upon acceptance are published on the portal [12].

3   Distributed CIML Resources

Resources in computational intelligence and machine learning are ubiquitously present in the internet. Countless webpages describing related topics, data repositories, individual authors' pages, computational tools, and others are available. A simple query "machine learning" returns about 1,760,000 "hits" in google.com search, and over 400,000 from bing.com. Similarly, the term "computational intelligence" is found 286,000 and over 8 million times on the two search services, respectively. Moreover, available online directories of resources provide no quality assurance. For example, the machine learning category in Google's directory has 260 websites, but with widely varying quality. This enormous amount of possible sites to look for information makes it difficult for potential users to access high quality resources, and discourages them from using CIML at all.

Figure 1. The main page of CIML virtual community portal

Automated discovery, reasoning and improved organization of distributed CIML resources can be seen as a first remedy of this situation. It requires careful evaluation of available resources. A standard classification of resources is needed before such an evaluation is attempted. Several taxonomies of machine learning methods are available in the literature [1, 3], but are not prepared from the perspective of potential users.

The section below describes the types of CIML-related resources that potential users may be interested in, including a brief classification of methods, their types and representation used. Then, typical application domains are presented, along with example application problems followed by how resources can be linked together. Each CIML resource available on the internet can be categorized based on the categories outlined here.

3.1 Classification of CIML Resources

Almost all published knowledge, results of studies, datasets, and other materials are available or can be made available on the internet. From the users'

perspective these materials can be classified into the main categories outlined below.

Web portals offer a multitude of resources in one place, usually organized and having a coherent graphical interface. Web portals may include some or all of the other types of resources described below. Several web portals with a scope covering different aspects of CIML are maintained all over the world. There is, however, no coordination and integration of these efforts. Although a very large number of CIML-related websites exists in the internet, relatively few of them are web portals composed of several web pages and multiple resources. Some examples of popular CIML-related web portals are: AI Topics[1] maintained by the Association for Advancement of Artificial Intelligence; Pascal Network[2] a multi-organization European initiative whose aim is to support research and education in areas related to CIML; Encyclopedia of Computational Intelligence[3] which is a part of the ScholarPedia initiative whose goal is to bring together a collection of peer-reviewed articles describing different aspects of CI. Other initiatives or networks with aims similar to the CIML VO are the Knowledge Discovery Network of Excellence[4] and Knowledge Discovery in Ubiquitous Environments[5]. An example of web portal directed toward support for end-users of machine learning solutions is the MiningMart project[6] that aims to build a library of data mining cases that can be reused for other projects.

Professional organizations are groups of people associated in order to pursue common professional goals. Multiple organizations whose missions are related to CIML exist, including the Association for Advancement of Artificial Intelligence (AAAI), the IEEE Computational Intelligence Society, and the International Machine Learning Society. Many of these CIML-related organizations maintain their often separate websites and web portals.

Educational and research institutions include universities, departments, research institutes and other organizations engaged in research and educational activities related to CIML.

Software and tools are computer programs which are either computer implementations of CIML methods or are auxiliary programs that help in applying such methods (i.e. data preprocessing). An important part of CIML research is of the applied nature and is developed with certain applications in mind. However, a significant portion of software developed for research purposes is never used by other researchers or end-users. There are several possible reasons for this situation: research software is rarely published, it is rarely properly documented and, finally, it can be hard to use. The quality of freely

---

1 http://www.aaai.org/AITopics/pmwiki/pmwiki.php/AITopics/HomePage

2 http://www.pascal-network.org

3 http://www.scholarpedia.org/article/Encyclopedia of computational intelligence

4 KDNet; http://www.kdnet.org

5 KDUbiq; http://www.kdubiq.org

6 http://mmart.cs.uni-dortmund.de

available shareware is uneven and hard to check before it is downloaded and used, despite often available user opinions. The qualifications of users writing opinions cannot be easily verified, unlike of those who peer review journal or conference articles.

The majority of research software available in the internet is maintained by individual researchers, laboratories or departments on their websites. For example, Weka [10] is a popular program that implements many machine learning methods in Java. The system is maintained by its creators at the University of Waikato[7] and many of the modules are contributed by researchers and programmers from all over the world. One initiative that aims at publishing software is organized by the Journal of Machine Learning Research[8], which, in addition to regular research papers, also considers software-related submissions.

Datasets are important resources in CIML that can be used for testing newly developed methods. Different datasets exist for testing CIML software designed to perform typical tasks such as learning classification models, learning predictive models, image recognition, and others.

Application problems are either standard benchmark problems used to test CIML methods, or currently unresolved or "open" areas to which researchers are trying to find solutions. Some problems, especially those used in learning and applying classification models may require datasets, and others, such as those used to test optimization methods on standard functions, are formulated in different ways. For example, testing evolutionary optimization methods on constrained problems can be done using the functions described by Liang [7].

Journals & magazines are the most common traditional method of communication between scientists. The majority of publishers provide webpages where potential readers and authors can find information about their journals, subscription information, author guidelines, etc. Open access journals are becoming increasingly popular to provide free access to the published material. For example, the Journal of Machine Learning Research provides free access to publications along with a relatively fast submission and review process.

Conferences & meetings both in-person and online provide the possibility to discuss ideas, present work, and meet others interested in similar areas. All major conferences have a presence in the web, although for many conferences the web address changes annually as it usually includes the year, hosting institution, etc., like for the 27th International Conference on Machine Learning[9].

---

7 http://www.cs.waikato.ac.nz/ ml/weka/
8 http://jmlr.csail.mit.edu/
9 http://www.icml2010.org

Publications include peer-reviewed journals, magazines, and conference papers along with book chapters and technical reports (sometimes also reviewed).

Online articles describe specific methods, software and users' opinions related to CIML. Here, by online articles we mean individual articles describing CIML-related topics . For example, Wikipedia consists of many articles related to CIML where, in essence, one article is equivalent to one term in Wikipedia.

Books provide instructional and special materials intended for students, researchers and practitioners. Some books are available online with free or paid access.

Repositories (software, datasets, problems) are collections of resources or links to resources of a specific type. For example, the most well known repository of machine learning datasets is hosted by University of California at Irvine[10]. Other repositories do not directly host the actual resources, but only list them. This is most often the case for publication databases such as DBLP, the best well known list of publications in computer science[11].

Blogs are websites with comments, observations, progress in performing tasks, etc., presented in reversed chronological order (usually stamped by the date of entry). This type of websites is often used by individuals who wish to share their experience, i.e. in using or developing software. An example of a machine learning related blog is Machine Learning Thoughts created by Olivier Bousquet[12].

Educational materials include posted recorded lectures, course webpages and syllabi, exercises, demos and everything else that helps students' learning. For example, a good collection of recorded CIML-related lectures is available at videolectures.net.

People are represented in the internet by their personal or institutional webpages.

Many resources available in the internet can be classified to more than one category. For example, the resources on the Weka website, although organized around the multipurpose machine learning software, consist of several additional types of resources: presentation of a book describing the software, publications related to the project, links to related webpages, manuals, information for potential developers etc.

---

10 http://archive.ics.uci.edu/ml/
11 http://www.informatik.uni-trier.de/ ley/db/
12 http://ml.typepad.com/

3.2 Classification of CIML Problems

Each of the CIML resources whose main categories are listed above is associated with one or more types of problems to be solved or tasks to be performed. General classification of problems with typically applied methods of solving them is depicted in Figure 2. In general, these problems can be reduced to the creation or application of models. A large portion of available software, published materials and methods relate to both categories, despite their clearly opposing character. For example, almost all rule learning programs have the ability to apply induced rules to classify new examples. In contrast, the goal of mining for association rules is to derive knowledge useful to analysts and decision makers, not necessarily to apply the rules to make predictions.
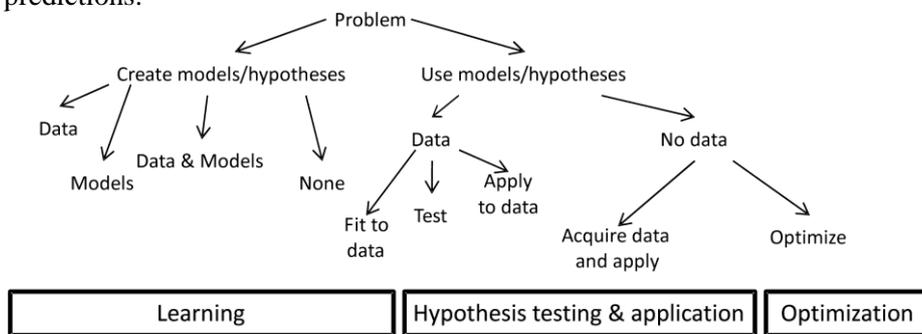
Fig. 2. Classification of CIML methods by functions/tasks performed

3.3 Application Domains

Applications of computational intelligence and machine learning methods are ubiquitously present in virtually all domains of science and engineering. Among the most popular application areas of CIML are healthcare, medicine, engineering, security, logistics and manufacturing.

3.4 Linking Resources Together

Resources do not exist alone. One of the main drawbacks of the currently available resources in the internet is the lack of their connectivity. It is hard to find related resources without going to basic web search tools. Consider the following scenario: A pediatrician, interested in improving outcomes for his patients with asthma is collecting data describing different patients' characteristics. He suspects that there is a strong correlation between some of the measured parameters and reflux respiration with cough. Being able to predict

cough from the collected data is important because it helps to avoid performing invasive tests which are often administered. To solve this problem he starts reading about existing applicable methods in concept learning. A webpage that describes concept learning is linked to available software, tutorials on how to use that software, recorded lectures about concept learning, and so on. Based on ratings and reviews, description, and properties, he chooses to use a rule learning program. The program is accompanied by additional resources: user's guide, tutorial, example datasets, publications describing details of the method, list of known users and researchers who can provide support (all of which can be on different internet sites), links to more general rule learning resources, other programs with similar capabilities, and so on. Having all of these resources available, the pediatrician can concentrate on solving the problem, not on struggling with software.

Most websites include hyperlinks to other resources. These links are created by websites' authors and link to their homepages, institutional websites, and some other resources that authors consider important. This mechanism is not sufficient as the creation of links is based on authors' judgment and bias (i.e., it is rare for researchers to provide links related to software created by competing groups).

The solution to linking together resources is based on discovered relationships between them. Every two related resources are linked together. Each link may be accompanied by its type, whenever that type is not implied by the types of resources. For example, when software is linked to a tutorial, it is implied that the tutorial is on how to use the software. On the other hand, a link type should be added whenever possible to provide additional information. For example, when a publication is linked to software, that publication can either describe the software, a theory behind the used method, or an application of that software to a problem. In the latter case, that publication would also be linked to the problem if it is present as a resource in the VO.

4 Search and Reasoning

4.1 Initial Resources Database

In the first stage of creating the initial resource database approximately 300 entries were collected, reviewed, and described. To discover the available resources, web search engines were used with carefully devised search phrases. Then, links from the known resources were followed. Used search phrases included e.g. "machine learning software", "computational intelligence tool", "rule learning methodology", etc. In order to find pages describing downloadable software, words like "download", "install", "manual" were added. For other types of resources adequate combinations of search words were applied.

The initial database was collected by a single person with an advanced level of expertise in CIML. The selection process put preference on the web sites that were well maintained (e.g. lack of broken links, recent updates, all functionality working well), easy-to-navigate, comprehensible (easy-to under-stand, complete and correct descriptions) and important (recognized in respective research areas). Even with these preferences, however, the quality of discovered resources varied depending on the methodological approach and application domain.

Resources are represented in a relational database containing their detailed descriptions and categorization, application domains, as well as categories for the research methods used. Details about the resources include their name, link to the website, general description, computer platforms and programming languages in case of software resources, expected level of the expertise in the area, and also tags related to intended audience. Moreover each resource was evaluated by the collector in terms of its quality, both related to the ease of use and overall.

Figure 3 shows a sample of a few entries after querying the database and transforming the results into more human-readable form. Each resource is presented with its name (which is a clickable link leading to the corresponding website), its type (e.g. platform, framework, collection, denote different level of integration between the elements), quality indicators (with the use of grey shading – lighter color denotes higher quality), indicators of the category of the methods (at present only 8 high-level categories are used, if a category applies, its 3-letter abbreviation has colored background), description, list of tags and software platform and programming language, if applicable. Quite often the entered text had to be abbreviated on the figure.

Currently the collected database of resources is undergoing the peer-review and error inspection process, and is being extended with additional resources.

4.2 Building Resources Database

Finding, describing, ranking, and linking together such a variety of distributed CIML resources is not a trivial task. Standard search techniques and engines that are designed to be multipurpose and serve multiple users have been found only partially helpful for the task. While the problem of intelligent search for information relevant to a specific user is widely studied [4], no specific search method fits directly into needs of collecting resources for the CIML VO.

Fig. 3. Sample entries from the web page presenting resources from the database

The starting point for search and evaluation of CIML resources is the manually constructed database briefly described above. Based on the database new resources are found by:

- **Searching for similar webpages or keywords**. This can be done using existing web search engines. Results of searches are based on keywords, thus results return not necessarily CIML relevant webpages.
- **Manually adding new resources.** Resources may be manually added to the list by VO's members, users, and developers.

- **Following hyperlinks on the existing webpages.** This technique is used by web crawlers used by internet search engines. Links on webpages related to CIML are likely to point to other relevant webpages, but not all of them.

After new resources have been found they need to be described using attributes, some of which were outlined in the previous section, and reviewed. While the process of deriving attributes can be partially automated, the review process needs to be done by CIML VO members qualified for the review. Depending on the resource type reviews are performed similarly to those done for scientific journals and conferences. The process may provide limited feedback to the websites' authors. A more comprehensive review process that ensures the quality of the resources would open the possibility to make them citeable.

4.3 Reasoning

The goal of reasoning is to provide the most relevant resources to users in the shortest possible time (to minimize search for that resources). Current and potential users of CIML have different experience, different levels of computer literacy, different expertise in application domains, and most importantly differing needs to solve different problems. A virtual organization intended for a wide range of users need to adapt to their needs and qualifications – resources should be useful to a computer science PhD student working on a dissertation related to machine learning, compared to a medical doctor who wants to analyze a dataset, or to an undergraduate student working on a CIML-related project.

Reasoning about resources is based on a dynamically created model that is based on users' interactions with the system. Each encounter is recorded in terms of accessed resources, selected preferences, etc. and can be used to induce the preference model. The model describes relationships between users' needs, available resources, and available experts (i.e., use CIML to build CIML collaboration). Such relationships take the form:

$$< p_1,...,p_k,u_1,...,u_l > \Rightarrow < r_1,...,r_m > \qquad (1)$$

in which p_i are the problem characteristics, u_i are user's characteristics (based on user profiles), and r_i are desired resource characteristics based on available CIML methods taxonomies such as presented in Section 3. The approach taken here is somewhat similar to meta-learning [2, 5, 10], but is not limited to the selection of methods or best parameters for a given problem, as is usually done in meta-learning. The relationship (1) can be modeled using multi-head attributional rules [6] in the form (2) that is a natural generalization of (1).

16

The rules are incrementally derived from data collected during users' interactions with the system.

$$RESOURCES \Leftarrow PREFERENCES \mathbin{|\_} EXCEPTIONS \qquad (2)$$

Here *RESOURCES* is a conjunction of characteristics of resources in the form of attributional conditions, *PREFERENCES* is a conjunction of a user's preferences and characteristics, and *EXCEPTIONS* is a conjunction of a user's preferences and characteristics that constitute exceptions to those included in *PREFERENCES*, or an explicit list of cases that are exceptions to the rule. Here is an example of such a rule:

```
[Input format = Excel or CSV] & [Representation =
Bayesian or rule-based]<== [Domain = pediatrics] &
[Data type = multitype or numeric]&
[Transparency=medium .. high] & [Data source =
patients' records]|_ [Relational data = yes] &
[De-individualized = no]
```

The rule can be paraphrased: *the input data format is excel or CSV and representation is Bayesian or rule-based if the domain is pediatrics, the data is multitype or numeric, and required transparency is between medium and high, except for when the data is in relational form, and it is not de-individualized.*

Implementation of this approach can be based on an extension of the AQ21 rule learning system [11]. The rationale for selecting this system is that it is a highly flexible and configurable rule learner able to handle multitype data, learn rules with exceptions, make use of user-defined quality criteria, automatically improve representation space, take advantage of background knowledge, and others. The preference rules to be learned will be in the form (2), which is a special form of rules that can be learned by AQ21. It is important to note that the form of rules used here is more general than standard classification rules induced by most programs, thus more complicated concepts can be captured.

The advising process should be able to minimize the number of questions that a user needs to answer in order to find relevant resources. We anticipate that no more than 6 questions will be necessary. To achieve this goal, a modified version the dynamic recognition method [9] that adjusts asked questions based on the initial problem statement and answers already provided. After analyzing the initial problem statement, the method selects a set of rules that potentially apply to the specific situation. Then, the program asks the question having the strongest presence in the *PREFERENCES* part of the rules. When an answer is provided, the set of rules is revised, and the process is repeated

until an answer or a set of equally probable answers is obtained. Note that the process may not lead to a unique selection of specific resources, but to a group of resources with desired characteristics. In fact, the system will initially give imprecise answers and improve precision when more specific rules are learned based on interaction with users.

5 Conclusion

Computational intelligence and machine learning (CIML) resources are ubiquitously present in the internet. These resources are, however, not categorized and their quality varies from site to site. CIML Virtual Organization aims at unifying resources, and providing high quality material to interested scientists, educators, and general public. To achieve this goal search and reasoning methodology based on CIML itself can be used. In the search phase new resources are identified and relationships between them are discovered. In the reasoning phase, resources are selected that best fit users' needs, expertise, application area, and qualifications.

References

1. Atanassov, K., Hadjiiski, M., 2008: *Generalized nets as tools for modeling of intelligent systems.* Proc. 4th Int. IEEE Conf. on Intelligent Systems, IIS '08, Varna, Bulgaria.
2. Brazdil, P.B. and Soares C., 2000: *Ranking Classification Algorithms Based on Relevant Performance Information.* Proc. 11th European Conf. on Machine Learning, Workshop on Meta-Learning: Building Automatic Advice Strategies for Model Selection and Method Combination, Barcelona, Spain.
3. Carbonell, T.J., Michalski, R.S. and Mitchell, T.M., 1983: *An Overview of Machine Learning. In Machine Learning: An Artificial Intelligence Approach*, R.S. Michalski, T.J. Carbonell and T.M. Mitchell (Eds.), TIOGA Publishing Co., Palo Alto, pp. 3-23.
4. Coyle, M. and Smyth B., October 2007: *Supporting intelligentWeb search.* ACM Trans. on Internet Technology, 7, 4.
5. Gagliolo, M. and Schmidhuber, J, 2006: Dynamic algorithm portfolios. Annals of Mathematics and Artificial Intelligence.
6. Glowinski, C. and Michalski, R.S., 2001 June: *Discovering Multi-head Attributional Rules in Large Databases*. 10th Int. Symposium on Intelligent Information Systems, Zakopane, Poland.
7. Liang, J.J., Runarsson, T.P., Mezura-Montes, E., Clerc, M., Suganthan, P.N., Coello Coello, C.A., and Deb, K., 2005: *Problem Definitions and Evaluation Criteria*. CEC 2006 Special Session on Constrained Real-Parameter Optimization, Tech Rep., Nanyang Technical University, Singapore.

8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009: *The WEKA Data Mining Software: An Update*. SIGKDD Explorations, Volume 11, Issue 1.

9. Michalski, R.S., 1986: *Dynamic Recognition: An Outline of Theory of How to Recognize Concepts without Matching Rules.* Rep. of the Intelligent Systems Group, ISG 86-16, UIUCDCS-F-86-965, Urbana.

10. Vilalta, R. and Drissi, Y., 2002: *A perspective view and survey of meta-learning.* Artificial Intelligence Review, 18, 2, pp. 77-95.

11. Wojtusiak, J., Michalski, R. S., Kaufman, K. and Pietrzykowski, J., November 13-15, 2006: *The AQ21 Natural Induction Program for Pattern Discovery: Initial Version and its Novel Features.* Proc. of The 18th IEEE Int. Conf. on Tools with Artificial Intelligence, Washington D.C.

12. Zurada, J.M., Mazurowski, M.A., Abdullin, A., Ragade, R.,Wojtusiak, J. and Gentle, J.E., 2009: *Building Virtual Community in Computational Intelligence and Machine Learning*, Computational Intelligence Magazine, 4, 1, pp. 43-54.