Machine Learning in the Life Sciences

How it is Used on a Wide Variety of Medical Problems and Data

KRZYSZTOF J. CIOS, LUKASZ A. KURGAN, AND MAREK REFORMAT

Over the years several definitions of machine learning have been proposed. One of the earliest ones read "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [1]. Another stated that machine learning is the "Ability of a computer program to generate a new data structure that is different than an old one, like production if…then…rules from input numerical or nominal data" [2], and a very broad definition simply specified "Things learn when they change their behavior in a way that makes them perform better in the future" [3]. The unifying theme of machine learning is that it is concerned with development of techniques that help extract knowledge/information from training data in an automatic way in order to discover some regularities and use them to build a general and accurate model able to make predictions for unseen data.

Life sciences, including biology and medicine, are a growing application area of machine learning. Medicine is largely an evidence-driven discipline where large quantities of relatively high-quality data are collected and stored in databases. The medical data are highly heterogeneous and are stored in numerical, text, image, sound, and video formats. They include clinical data (symptoms, demographics, biochemical tests, diagnoses and various imaging, video, vital signals, etc.), logistics data (charges and costs, policies, guidelines, clinical trials, etc.), bibliographical data, and molecular data. Bioinformatics, which concerns the latter type of data, conceptualizes biology in terms of molecules and applies "informatics" techniques, derived from disciplines such as applied mathematics, computer science, and statistics to understand and organize the information associated with these molecules on a large scale [4]. In other words, bioinformatics encompasses analysis of molecular data expressed in the form of nucleotides, amino acids, DNA, RNA, peptides, and proteins. The sheer amount and breadth of data requires development of efficient methods for knowledge/information extraction that can cope with the size and complexity of the accumulated data. There are numerous examples of successful applications of machine learning (ML) in areas of *diagnosis* and *prevention* [5]–[7], *prognosis* and *therapeutic decision making* [8], [9]. ML algorithms are used for *discovering* new diseases [10], *finding* predictive and therapeutic biomarkers [11], and *detecting* relationships and structure among the clinical data [12]–[14]. ML contributes to the enhancement of *management* and *information retrieval* processes leading to development of intelligent (involving ontologies and natural language processing) and integrated (across repositories) literature searches [15], [16]. ML techniques are also used to *modify medical procedures* in order to reduce cost and improve perceived patient's experience and outcomes [17], [18].

This special issue presents contributions chosen from a special session on *Applications of Machine Learning in Medicine and Biology* at the 4th International Conference on Machine Learning and Applications, held in December 2005 in Los Angeles. It includes extended versions of seven articles chosen from the seventeen presented at the special session on "Applications of Machine Learning in Medicine and Biology" and two invited articles. Although some surveys show increasing interest in applications of machine learning in bioinformatics [19], [20], this special issue demonstrates that machine learning is also successfully used on a wide variety of medical problems and data.

A breadth of applications and tools are presented that range from bioinformatics (microarray analysis, chromosome and proteome databases, modeling of inhibition of metabolic networks), through signal analysis (echocardiograph images and electroencephalograph time series), drug delivery, information retrieval, to software for pattern recognition in biomedical data. The authors use a variety of techniques such as association rules, feature selection, Fisher's linear discriminant analysis, inductive logic programming, linear auto regression, neural networks, and reinforcement learning to achieve their goals. To compare how ML techniques are used in medical problems, the articles are organized according to a data mining and knowledge discover process model, which is used to guide ML driven projects in biology and medicine [14], [21]–[24]. The model describes and organizes all activities of a ML project into a sequence of six steps [24], [25]:

- understanding the problem, where authors present project goals, current solutions and domain terminology, and translate the medical problem into the ML domain
- understanding the data, where the corresponding medical data is described and analyzed with respect to the underlying ML problem

 0739-5175/07/$25.00©2007IEEE

- ➤ preparation of the data, in which the data preprocessing methods are applied.
- ➤ data mining, in which the prepared data is processed with ML techniques.
- ➤ evaluation of the discovered knowledge, where the results provided in the previous step are evaluated
- ➤ using the discovered knowledge, in which the authors describe how the generated knowledge is deployed.

The nine articles included in this special issue are divided into three groups. The first four articles are *general applications* in medicine:

1) The cDNA microarrays technology enables the screening of a biological sample for expressions of thousands of genes simultaneously. Among all the genes assayed, only a small fraction of them actually participate in the biological process of interest. In the article titled "New Criteria for Selecting Differentially Expressed Genes," Loo, Roberts, Hrebien, and Kam propose two new data mining criteria for the selection of these differentially expressed genes.

2) Drug dosing in chronic conditions often has a form of recurrent trial and error control process. The article by Gaweda, Muezzinoglu, Aronoff, Jacobs, Zurada, and Brier titled "Using Clinical Information in Goal-Oriented Learning" presents a numerical framework based on the paradigm of reinforcement learning, which mathematically formalizes this process and enables computer-supported individualized drug administration.

3) In "Modeling the Effects of Toxins in Metabolic Networks," Tamaddoni-Nezhad, Chaleil, Kakas, Sternberg, Nicholson, and Muggleton use a logic-based representation and a combination of abduction and induction to model inhibition in metabolic networks. Inhibition is very important from the therapeutic point of view since many substances designed to be used as drugs can have an inhibitory effect on other enzymes. Any system able to predict the inhibitory effect of substances on the metabolic network would therefore be very useful in assessing the potential harmful side-effects of drugs. According to the domain experts, one of the hypothesized enzymes (i.e., EC2.6.1.39) has been already known to be inhibited by hydrazine. Another hypothesis suggested by the model agrees with the speculations about the inhibition of enzyme EC4.3.2.1 by hydrazine. Experimental evaluations in vivo are required to test this hypothesis.

4) In the article titled "Multilabel Associative Classification of MEDLINE Articles into MeSH Keywords," Rak, Kurgan, and Reformat propose an automated method for classification of medical articles into the structure of document repositories, which aims at supporting currently performed extensive manual work. The proposed method classifies articles from the largest medical repository, MEDLINE, using associative classification, a state-of-the-art data mining technique. The method considers re-occurring features of articles and, most importantly, multilabel characteristic of the MEDLINE data. The results of experiments performed using several different classification approaches are compared, and pros and cons of different measures of classification quality are discussed. The results show high potential of the method to support tedious work associated with maintaining large databases of medical documents.

The next three articles focus on medical signal analysis:

5) "Automated Heart Abnormality Detection Using Sparse Linear Classifiers" article by Qazi, Fung, Krishnan, Bi, Rao, and Katz describes a fully automated and robust technique for detection of diseased hearts based on detection and tracking of the endocardium and epicardium of the left ventricle. The authors used a novel feature selection and classification technique based on mathematical programming to obtain classifiers that depend only on a small subset of numerical features extracted from dual-contours tracked through time. They verified the quality of the proposed system on echocardiograms collected in routine clinical practice using the cross-validation tests and a held-out set of unseen echocardiography images.

6) Rezek, Roberts, and Conradt in their article titled "Increasing the Depth of Anesthesia Assessment" proposed a model that generalizes a class of polyspectral models. The authors show that the model estimation can be done in the Bayesian framework and that it requires less data than the traditional estimation methods. They test the model on several electroencephalographic signals recorded during exposure to different anaesthetic agents and indicate that polyspectra contains information beyond the standard spectra, which helps discriminate between wake and anaesthetised states in two out of three anaesthetic agents or agent combinations.

7) A method for the reliable detection of epileptic seizures in electroencephalographic data using radial basis function neural networks combined with one of several preprocessing methods is presented in the article "Epileptic Seizure Detection" by Schuyler, White, Staley, and Cios. The article evaluates several preprocessing methods including Fourier and wavelet transforms. The possibility of using this method for seizure prediction is also investigated by the authors.

The last two articles describe bioinformatics tools. They include a software platform and a database, which ML researchers developed for the medical community.

8) "A Pattern Recognition Application Framework for Biomedical Datasets" article by Vivanco, Demko, Jarmasz, Somorjai, and Pizzi describes a multi-platform (Linux, Windows, Mac OS) open-source C++ application framework for the analysis and visualization of biomedical datasets. This software tool takes a full advantage of MPI for cluster computing, and is currently being enhanced with an easy to program agent architecture for distributed computing. It has a plug-in architecture to facilitate the coupling of third party libraries and can be integrated with MATLAB via an efficient data sharing mechanism.

9) Nguyen, Thaicharoen, Lacroix, Gardiner, and Cios present a chromosome 21 (chr21) database in their "A Comprehensive Human Chromosome 21 Database" article. Their goal is to store all chr21-related gene and protein information. The authors designed an easy-to-use user interface, called GeneQuest, which enables even inexperienced users to fully utilize the database in an efficient manner.The database embraces a wide range of information including chr21 gene structures, protein post-translational modifications and interactions, chr21 orthologs in model organisms and their phenotypes of RNAi, and their protein-protein interactions. They also added a predictor of a protein-protein interaction function that is based on Markov random fields method. The database and its associated tools can be accessed at http://chr21db.cudenver.edu.

We hope that the *IEEE EMB Magazine* readers will enjoy this special issue and benefit from the wealth of information conveyed by the authors in their articles.

Acknowledgments

We extend our thanks to the reviewers who have made this special issue possible: Charles Anderson (Colorado State University), John Anderson (University of Manitoba), Grzegorz Boratyn (University of Louisville), Abdennour El Rhalibi (Liverpool John Moores University), Dragan Gamberger (Rudjer Boskovic Institute), Altay Guvenir (Bilkent University), Rumen Kountchev (Technical University of Sofia), James Tin-Yau KWOK (Hong Kong University of Science and Technology), Dunja Mladenic (Jozef Stefan Institute), Daniel Neagu (University of Bradford), Vasile Palade (Oxford University), Witold Pedrycz (University of Alberta), Yonghong Peng (University of Bradford), Petra Perner (Institute of Computer Vision and Applied Computer Sciences), Petia Radeva (Universitat Autònoma de Barcelona), Jagath Rajapakse (Nanyang Technological University), Bernardete Ribeiro (University of Coimbra), Jungpil Shin (University of Aizu), Tomasz Smolinski (University of Louisville), Ryszard Tadeusiewicz (AGH University of Science and Technology), Gheorghe Tecuci (George Mason University), Du Zhang (California State University, Sacramento) and Jozef Zurada (University of Louisville).

Krzysztof J. Cios received the M.S. and Ph.D. degrees from the AGH University of Science and Technology, Krakow, the MBA degree from the University of Toledo, Ohio, and the D.Sc. degree from the Polish Academy of Sciences. He is currently a professor at the University of Colorado at Denver and Health Sciences Center, and Associate Director of the University of Colorado Bioenergetics Institute. He directs Data Mining and Bioinformatics Laboratory. Dr. Cios is a well-known researcher in the areas of learning algorithms, biomedical informatics and data mining. NASA, NSF, American Heart Association, Ohio Aerospace Institute, NATO, US Air Force and NIH have funded his research. He published three books, about 150 journal and conference articles and 12 book chapters, and has edited five special issues of journals. Dr. Cios is a Foreign Member of the Polish Academy of Arts and Sciences.

Lukasz A. Kurgan received his M.Sc. with honors (and was recognized with an Outstanding Student Award) in automation and robotics from the AGH University of Science and Technology, Krakow, in 1999, and his Ph.D. in computer science from the University of Colorado at Boulder, in 2003. He is currently an assistant professor in the Department of Electrical and Computer Engineering at the University of Alberta in Edmonton, Alberta, Canada. His research interests include data mining and knowledge discovery, machine learning, computational biology, and bioinformatics. He currently serves as an associate editor of *Neurocomputing*. Dr. Kurgan is a member of the IEEE, ACM, and ISCB.

Marek Reformat received his M.Sc. (with honors) from Technical University of Poznan, Poland, and his Ph.D. from University of Manitoba, Winnipeg, Manitoba, Canada. Currently, he is with the Department of Electrical and Computer Engineering at the University of Alberta, Edmonton, Alberta, Canada. Dr. Reformat has been a member of program committees of several conferences related to computational intelligence and evolutionary computing. He is a member of the IEEE and ACM.

References

[1] T. Mitchell, *Machine Learning*, New York:McGraw-Hill, 1997

[2] K. Cios, W. Pedrycz, and R. Swiniarski, *Data Mining Methods for Knowledge Discovery*, Norwell, MA: Kluwer, 1998.

[3] I.H. Witten, and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, San Mateo, CA: Morgan Kaufmann, 2000.

[4] N.M. Luscombe, D. Greenbaum, and M., Gerstein, "What is bioinformatics? A proposed definition and overview of the field," *Methods Inform. Medi.*, vol. 40, no. 4, pp. 346–58, 2001.

[5] H. Guo and A.K. Nandi, "Breast cancer diagnosis using genetic programming generated feature," *Pattern Recognit.*, vol. 39, no. 5, pp. 980–987, 2006.

[6] H. Yan, Y. Jiang, J. Zheng, C. Peng, and Q. Li, "A multilayer perceptron-based medical decision support system for heart disease diagnosis," *Expert Syst. Applicat.*, vol. 30, no. 2, pp.272–281, 2006.

[7] H. Shin and M.K. Markey, "A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples," *J. Biomed. Informatics*, vol. 39, no. 2, pp. 227–248, 2006.

[8] S.S. Thorgeirsson, J.-S. Lee, and J.W. Grisham, "Molecular prognostication of liver cancer: End of the beginning," *J. Hepatology*, vol. 44, no. 4, pp. 798–805, 2006.

[9] Z. Gao, G. Tomaselli, C. Wei, and R. Winslow, "Constructing gene expression-based diagnostic rules for understanding individualized etiology of heart failure," *J. Cardiothoracic-Renal Res.*, vol. 1, no. 1, pp. 33–40, 2006.

[10] C. Baumgartner, G. Mátyás, B. Steinmann, M. Eberle, J.I. Stein, and D. Baumgartner, "A bioinformatics framework for genotype-phenotype correlation in humans with Marfan syndrome caused by FBN1 gene mutations," *J. Biomed. Informatics*, vol. 39, no. 2, pp. 171–183, 2006.

[11] C. Baumgartner and D. Baumgartner, "Biomarker discovery, disease classification, and similarity query processing on high-throughput MS/MS data of inborn errors of metabolism," *J. Biomolecular Screening*, vol. 11, no. 1, pp. 90–99, 2006.

[12] S.P. Linke, T.M. Bremer, C.D. Herold, G. Sauter, and C. Diamond, "A multimarker model to predict outcome in tamoxifen-treated breast cancer patients," *Clinical Cancer Res.*, vol. 12, no. 4, pp. 1175–1183, 2006.

[13] L. Ramirez, N.G. Durdle, V.J. Raso, and D.L. Hill, "A support vector machines classifier to assess the severity of idiopathic scoliosis from surface topography," *IEEE Trans. Inform. Technol. Biomed.*, vol. 10, no. 1, pp. 84–91, 2006.

[14] L. Kurgan, K. Cios, M. Sontag, and F. Accurso, "Mining the cystic fibrosis data," in *Next Generation of Data-Mining Applications,* J. Zurada and M. Kantardzic, Eds. Piscataway, NJ: IEEE Press and New York: Wiley, 2005, pp. 415–444.

[15] E.V. Bernstam, J.R. Herskovic, Y. Aphinyanaphongs, C.F. Aliferis, M.G. Sriram, and W.R. Hersh, "Using citation data to improve retrieval from MEDLINE," *J. Amer. Medical Informatics Assoc.*, vol. 13, no. 1, pp. 96–105, 2006.

[16] Z.-Z. Hu, I. Mani, V. Hermoso, H. Liu, and C.H. Wu, "IProLINK: An integrated protein resource for literature mining," *Comput. Biol. Chem.*, vol. 28, no. 5–6, pp. 409–416, 2004.

[17] S.-K. Ng, G.J. McLachlan, and A.H. Lee, "An incremental EM-based learning approach for on-line prediction of hospital resource utilization," *Artificial Intelligence in Medicine*, vol. 36, no. 3, pp. 257–267, 2006.

[18] K. Morik, M. Imboff, P. Brockhausen, T. Joachims, and U. Gather, "Knowledge discovery and knowledge validation in intensive care," *Artif. Intell. Med.*, vol. 19, no. 3, pp. 225–249, 2000.

[19] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J.A. Lozano, R. Armañanzas, G. Santafé, A. Perez, and V. Robles, "Machine learning in bioinformatics," *Briefings Bioinformat.*, vol. 7, no. 1, pp. 86–112, 2006.

[20] H. Bhaskar, D.C. Hoyle, and S. Singh, "Machine learning in bioinformatics: A brief survey and recommendations for practitioners, *Comput. Biol. Med.*, vol. 36, no. 1, pp. 1106-1125, 2006.

[21] L. Maruster, T. Weijters, G. De Vries, A. Van den Bosch, and W. Daelemans, "Logistic-based patient grouping for multi-disciplinary treatment," *Artif. Intell. Med.*, vol. 26, no. 1–2, pp. 87–107, 2002.

[22] S. Ganzert, J. Guttmann, K. Kersting, R. Kuhlen, C. Putensen, M. Sydow, and S. Kramer, "Analysis of respiratory pressure-volume curves in intensive care medicine using inductive machine learning," *Artif. Intell. Med.*, vol. 26, no. 1–2, pp. 69–86, 2002.

[23] P. Perner, H. Perner, and B. Mulle, "Mining knowledge for HEp-2 cell image classification," *Artif. Intell. Med.*, vol. 26, no. 1–2, pp. 161–173, 2002.

[24] K. Cios, A. Teresinska, S. Konieczna, J. Potocka, and S. Sharma, "Diagnosing myocardial perfusion from PECT bull's-eye maps–A knowledge discovery approach," *IEEE Eng. Med. Biol. Mag.*, vol. 19, no. 4, pp. 17–25, 2000.

[25] K. Cios and L. Kurgan, "Trends in data mining and knowledge discovery," in *Adv. Tech. Knowl. Discov.Data Mining,* N. Pal and L. Jain Eds. New York: Springer-Verlag, 2005, pp. 1–26.