Discovery of Process Models from Data and Domain Knowledge: A Rough-Granular Approach

Andrzej Skowron

Institute of Mathematics,
Warsaw University
Banacha 2, 02-097 Warsaw, Poland
skowron@mimuw.edu.pl

Extended Abstract

The rapid expansion of the Internet has resulted not only in the ever growing amount of data therein stored, but also in the burgeoning complexity of the concepts and phenomena pertaining to those data. This issue has been vividly compared [14] by the renowned statistician, prof. Friedman of Stanford University, to the advances in human mobility from the period of walking afoot to the era of jet travel. These essential changes in data have brought new challenges to the development of new data mining methods, especially that the treatment of these data increasingly involves complex processes that elude classic modeling paradigms. "Hot" datasets like biomedical, financial or netuser behavior data are just a few examples. Mining such temporal or stream data is on the agenda of many research centers and companies worldwide (see, e.g., [31, 1]). In the data mining community, there is a rapidly growing interest in developing methods for the discovery of structures of temporal processes from data. Works on discovering models for processes from data have recently been undertaken by many renowned centers worldwide (e.g., [34, 19, 36, 9], www.isle.org/~langley/, soc.web.cse.unsw.edu.au/bibliography/discovery/index.html).

We discuss a research direction for discovery od process models from data and domain knowledge within the program *Wisdom technology* (wistech) outlined recently in [15, 16].

Wisdom commonly means *rightly judging* based on available knowledge and interactions. This common notion can be refined. By *wisdom*, we understand an adaptive ability to make judgments correctly (in particular, correct decisions) to a satisfactory degree, having in mind real-life constraints. The intuitive nature of wisdom understood in this way can be metaphorically expressed by the so-called *wisdom equation* as shown in (1).

$$wisdom = adaptive\ judgment\ +\ knowledge\ +\ interaction. \tag{1}$$

Wisdom could be treated as a certain type of knowledge. Especially, this type of knowledge is important at the highest level of hierarchy of meta-reasoning in intelligent agents.

A. Ghosh, R.K. De, and S.K. Pal (Eds.): PReMI 2007, LNCS 4815, pp. 192–197, 2007.

Wistech is a collection of techniques aimed at the further advancement of technologies to acquire, represent, store, process, discover, communicate, and learn *wisdom* in designing and implementing intelligent systems. These techniques include approximate reasoning by agents or teams of agents about vague concepts concerning real-life, dynamically changing, usually distributed systems in which these agents are operating. Such systems consist of other autonomous agents operating in highly unpredictable environments and interacting with each others. Wistech can be treated as the successor of database technology, information management, and knowledge engineering technologies. Wistech is the combination of the technologies represented in equation (1) and offers an intuitive starting point for a variety of approaches to designing and implementing computational models for wistech in intelligent systems.

Knowledge technology in wistech is based on techniques for reasoning about knowledge, information, and data, techniques that enable to employ the current knowledge in problem solving. This includes, e.g., extracting relevant fragments of knowledge from knowledge networks for making decisions or reasoning by analogy.

Judgment technology in wistech is covering the representation of agent perception and adaptive judgment strategies based on results of perception of real life scenes in environments and their representations in the agent mind. The role of judgment is crucial, e.g., in adaptive planning relative to the Maslov Hierarchy of agents' needs or goals. Judgment also includes techniques used for perception, learning, analysis of perceived facts, and adaptive refinement of approximations of vague complex concepts (from different levels of concept hierarchies in real-life problem solving) applied in modeling interactions in dynamically changing environments (in which cooperating, communicating, and competing agents exist) under uncertain and insufficient knowledge or resources.

Interaction technology includes techniques for performing and monitoring actions by agents and environments. Techniques for planning and controlling actions are derived from a combination of judgment technology and interaction technology.

There are many ways to build foundations for wistech computational models. One of them is based on the *rough-granular computing* (RGC). Rough-granular computing (RGC) is an approach for constructive definition of computations over objects called granules, aiming at searching for solutions of problems which are specified using vague concepts. Granules are obtained in the process called granulation. Granulation can be viewed as a human way of achieving data compression and it plays a key role in implementing the divide-and-conquer strategy in human problem-solving [38]. The approach combines rough set methods with other soft computing methods, and methods based on granular computing (GC). RGC is used for developing one of the possible wistech foundations based on approximate reasoning about vague concepts.

As an opening point to the presentation of methods for discovery of process models from data we use the proposal by Zdzisław Pawlak. He proposed in 1992 [27] to use data tables (information systems) as specifications of concurrent

systems. Since then, several methods for synthesis of concurrent systems from data have been developed (see, e.g., [32]).

Recently, it became apparent that rough set methods and information granulation have set out a promising perspective to the development of approximate reasoning methods in multi-agent systems. At the same time, it was shown that there exist significant limitations to prevalent methods of mining emerging very large datasets that involve complex vague concepts, phenomena or processes (see, e.g., [10, 30, 35]). One of the essential weaknesses of those methods is the lack of ability to effectively induce the approximation of complex concepts, the realization of which calls for the discovery of highly elaborated data patterns. Intuitively speaking, these complex target concepts are too far apart from available low-level sensor measurements. This results in huge dimensions of the search space for relevant patterns, which renders existing discovery methods and technologies virtually ineffective. In recent years, there emerged an increasingly popular view (see, e.g., [12, 18]) that one of the main challenges in data mining is to develop methods integrating the pattern and concept discovery with domain knowledge.

In this lecture, the dynamics of complex processes is specified by means of vague concepts, expressed in natural languages, and of relations between those concepts. Approximation of such concepts requires a hierarchical modeling and approximation of concepts on subsequent levels in the hierarchy provided along with domain knowledge. Because of the complexity of the concepts and processes on top levels in the hierarchy, one can not assume that fully automatic construction of their models, or the discovery of data patterns required to approximate their components, would be straightforward. We propose to use in discovery of process models and their components through an interaction with domain experts. This interaction allows steering the discovery process, therefore makes it computationally feasible. Thus, the proposed approach transforms a data mining system into an experimental laboratory, in which the software system, aided by human experts, will attempt to discover: (i) process models from data bounded by domain constraints, (ii) patterns relevant to user, e.g., required in the approximation of vague components of those processes.

This research direction has been pursued by our team, in particular, toward the construction of classifiers for complex concepts (see, e.g., [2–4, 6–8, 11, 20–23]) aided by domain knowledge integration. Advances in recent years indicate a possible expansion of so far conducted research into discovery of models for processes from temporal or spatio-temporal data involving complex objects.

We discuss the rough-granular modeling (see, e.g., [29]) as the basis for discovery of processes from data. We also outline some perspectives of the presented approach for application in areas such as prediction from temporal financial data, gene expression networks, web mining, identification of behavioral patterns, planning, learning interaction (e.g., cooperation protocols or coalition formation), autonomous prediction and control by UAV, summarization of situation, or discovery of language for communication.

The novelty of the proposed approach for the discovery of process models from data and domain knowledge lies in combining, on one side, a number of novel methods of granular computing for wistech developed using the rough set methods and other known approaches to the approximation of vague, complex concepts (see, e.g., [2–8, 17, 20–23, 25, 26, 28, 29, 37, 38]), with, on the other side, the discovery of process' structures from data through an interactive collaboration with domain experts(s) (see, e.g., [2–8, 17, 20–23, 29]).

Acknowledgments

The research has been supported by the grant from Ministry of Scientific Research and Information Technology of the Republic of Poland.

Many thanks to Mr Tuan Trung Nguyen for suggesting many helpful ways to improve this article.

References

1. Aggarwal, C. (ed.): Data Streams: Models and Algorithms. Springer, Berlin (2007)
2. Bazan, J., Peters, J.F., Skowron, A.: Behavioral pattern identification through rough set modelling. In: Ślęzak, D., et al. (eds.) pp. 688–697 [33] (2005)
3. Bazan, J., Skowron, A.: On-line elimination of non-relevant parts of complex objects in behavioral pattern identification. In: Pal, S.K., et al. (eds.) pp. 720–725 [24](2005)
4. Bazan, J., Skowron, A.: Classifiers based on approximate reasoning schemes. In: Dunin-Kęplicz, B., et al. (eds.) pp. 191–202 [13] (2005)
5. Bazan, J., Skowron, A., Swiniarski, R.: Rough sets and vague concept approximation: From sample approximation to adaptive learning. In: Peters, J.F., Skowron, A., Grzymała-Busse, J.W., Kostek, B., Świniarski, R.W., Szczuka, M. (eds.) Transactions on Rough Sets I. LNCS, vol. 3100, pp. 39–63. Springer, Heidelberg (2004)
6. Bazan, J., Kruczek, P., Bazan-Socha, S., Skowron, A., Pietrzyk, J.J.: Risk pattern identification in the treatment of infants with respiratory failure through rough set modeling. In: Proceedings of IPMU 2006, Paris, France, Paris, July 2-7, 2006, pp. 2650–2657. Éditions E.D.K (2006)
7. Bazan, J., Kruczek, P., Bazan-Socha, S., Skowron, A., Pietrzyk, J.J.: Automatic planning of treatment of infants with respiratory failure through rough set modeling. In: Greco, S., Hata, Y., Hirano, S., Inuiguchi, M., Miyamoto, S., Nguyen, H.S., Słowiński, R. (eds.) RSCTC 2006. LNCS (LNAI), vol. 4259, pp. 418–427. Springer, Heidelberg (2006)
8. Bazan, J.: Rough sets and granular computing in behavioral pattern identification and planning. In: Pedrycz, W., et al. (eds.) [29] (2007) (in press)
9. Borrett, S.R., Bridewell, W., Langely, P., Arrigo, K.R.: A method for representing and developing process models. Ecological Complexity 4(1-2), 1–12 (2007)
10. Breiman, L.: Statistical modeling: The two Cultures. Statistical Science 16(3), 199–231 (2001)
11. Doherty, P., Łukaszewicz, W., Skowron, A., Szałas, A.: Knowledge Representation Techniques: A Rough Set Approach. Studies in Fuzziness and Soft Computing 202. Springer, Heidelberg (2006)

12. Domingos, P.: Toward knowledge-rich data mining. Data Mining and Knowledge Discovery 15, 21–28 (2007)
13. Dunin-Kęplicz, B., Jankowski, A., Skowron, A., Szczuka, M.: Monitoring, Security, and Rescue Tasks in Multiagent Systems (MSRAS 2004). Series in Soft Computing. Springer, Heidelberg (2005)
14. Friedman, J.H.: Data mining and statistics. What's the connection? Keynote Address. In: Proceedings of the 29th Symposium on the Interface: Computing Science and Statistics, Houston, Texas (May 1997)
15. Jankowski, A., Skowron, A.: A wistech paradigm for intelligent systems. In: Transactions on Rough Sets VI: Journal Subline. LNCS, vol. 4374, pp. 94–132. Springer, Heidelberg (2006)
16. Jankowski, A., Skowron, A.: Logic for artificial intelligence: The Rasiowa - Pawlak school perspective. In: Ehrenfeucht, A., Marek, V., Srebrny, M. (eds.) Andrzej Mostowski: Reflections on the Polish School of Logic, IOS Press, Amsterdam (2007)
17. Jankowski, A., Skowron, A.: Wisdom Granular Computing. In: Pedrycz, W., et al. (eds.) (in press 2007)
18. Kriegel, H.-P., Borgwardt, K.M., Kröger, P., Pryakhin, A., Schubert, M., Zimek, A.: Future trends in data mining. Data Mining and Knowledge Discovery 15(1), 87–97 (2007)
19. de Medeiros, A.K.A., Weijters, A.J.M.M., van der Aalst, W.M.P.: Genetic process mining: an experimental evaluation. Data Mining and Knowledge Discovery 14, 245–304 (2007)
20. Nguyen, H.S., Bazan, J., Skowron, A., Nguyen, S.H.: Layered learning for concept synthesis. In: Peters, J.F., Skowron, A., Grzymała-Busse, J.W., Kostek, B., Świniarski, R.W., Szczuka, M. (eds.) Transactions on Rough Sets I. LNCS, vol. 3100, pp. 187–208. Springer, Heidelberg (2004)
21. Nguyen, T.T.: Eliciting domain knowledge in handwritten digit recognition. In: Pal, S., et al. (eds.) pp. 762–767 [24] (2005)
22. Nguyen, T.T.: Outlier and exception analysis in rough sets and granular computing. In: Pedrycz, W., et al. (eds.) [29] (in press 2007)
23. Nguyen, T.T., Willis, C.P., Paddon, D.J., Nguyen, S.H., Nguyen, H.S.: Learning Sunspot Classification. Fundamenta Informaticae 72(1-3), 295–309 (2006)
24. Pal, S.K., Bandyopadhyay, S., Biswas, S. (eds.): PReMI 2005. LNCS, vol. 3776, pp. 18–22. Springer, Heidelberg (2005)
25. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
26. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning about Data. System Theory, Knowledge Engineering and Problem Solving, vol. 9. Kluwer Academic Publishers, The Netherlands, Dordrecht (1991)
27. Pawlak, Z.: Concurrent versus sequential the rough sets perspective. Bulletin of the EATCS 48, 178–190 (1992)
28. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences 177(1): 3–27; Rough sets: Some extensions. Information Sciences 177(1): 28–40; Rough sets and boolean reasoning. Information Sciences 177(1): 41–73 (2007)
29. Pedrycz, W., Skowron, A., Kreinovich, V. (eds.): Handbook of Granular Computing. John Wiley & Sons, New York (in press)
30. Poggio, T., Smale, S.: The mathematics of learning: Dealing with data. Notices of the AMS 50(5), 537–544 (2003)

31. Roddick, J.F., Hornsby, K., Spiliopoulou, M.: An updated bibliography of temporal, spatial and spatio- temporal data mining research. In: Roddick, J.F., Hornsby, K. (eds.) TSDM 2000. LNCS (LNAI), vol. 2007, Springer, Heidelberg (2001)
32. Suraj, Z.: Rough set methods for the synthesis and analysis of concurrent processes. In: Polkowski, L., Lin, T.Y., Tsumoto, S. (eds.) Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems, Studies in Fuzziness and Soft Computing, vol. 56, pp. 379–488. Springer, Heidelberg (2000)
33. Ślęzak, D., Yao, J., Peters, J.F., Ziarko, W., Hu, X.(eds.): RSFDGrC 2005. LNCS (LNAI), vol. 3642. Springer, Heidelberg (2005)
34. Unnikrishnan, K.P., Ramakrishnan, N., Sastry, P.S., Uthurusamy, R.: 4th KDD Workshop on Temporal Data Mining: Network Reconstruction from Dynamic Data Aug 20, 2006, The Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data (KDD 2006) August 20 - 23, 2006 Philadelphia, USA (2006), http://people.cs.vt.edu/ ramakris/kddtdm06/cfp.html
35. Vapnik, V.: Statistical Learning Theory. John Wiley & Sons, New York (1998)
36. Wu, F.-X.: Inference of gene regulatory networks and its validation. Current Bioinformatics 2(2), 139–144 (2007)
37. Zadeh, L.A.: A new direction in AI-toward a computational theory of perceptions. AI Magazine 22(1), 73–84 (2001)
38. Zadeh, L.A.: Generalized theory of uncertainty (GTU)-principal concepts and ideas. Computational Statistics and Data Analysis 51, 15–46 (2006)