Semantic data mining tutorial @ ECML/PKDD'2011

Introduction

The term **semantic data mining** denotes a data mining approach where domain ontologies are used as background knowledge. Such approach is motivated by large amounts of data that are increasingly becoming openly available and described using real-life ontologies represented in Semantic Web languages, arguably most extensively in the domain of biology. This recently opened up the possibility for interesting large-scale and real-world semantic applications.

The availability of semantically annotated data poses requirements for new kinds of approaches for data mining that would be able to deal with the complexity, and expressivity of the semantic representation languages, leverage on availability of ontologies and explicit semantics of the described resources, and account for novel assumptions (e.g., open world) that underlie reasoning services exploiting ontologies.

The tutorial will address the above issues, focusing on the problems of how machine learning techniques can work directly on the richly structured Semantic Web data, exploit ontologies, and the Semantic Web technologies, what is the value added of machine learning methods exploiting ontologies, and what are the challenges for developers of semantic data mining methods. It will also contain demonstrations of tools supporting semantic data mining.

Outline

The tutorial will present the topic of semantic data mining from three complementary perspectives.

Firstly, it will present a **general framework for semantic data mining**, following the work [NVTL09]. The first part of the tutorial will also discuss a new method for **semantic subgroup discovery**: g-SEGS. It will be accompanied with a presentation of the developed tool, a part of Orange4WS environment.

The second part of tutorial will cover the topic of **learning from description logics (DL-learning)**, motivated by the fact that the standard Web ontology language, OWL, is theoretically based on description logics. This will include a demo of a tool supporting DL-learning (a plugin to the Rapid Miner system).

Finally, the third part of the tutorial will cover the topic of **semantic meta-mining**. This approach has three features that distinguish it from its predecessors. First, more than in previous work, it adopts a process-oriented approach where meta-learning is applied to support design choices at different stages of the *complete data mining process or workflow*. Second, it complements dataset descriptions with an *in-depth analysis and characterization of algorithms*—their underlying assumptions, optimization goals and strategies, the models and patterns they generate. Finally, it *relies on a data mining ontology which distills extensive background knowledge concerning knowledge discovery itself*.

A more detailed outline is presented below (Download the whole set of slides):

Part I Introduction to semantic data mining (presenters: Nada Lavrac, Anze Vavpetic) Download slides Download demo

- Framework for semantic data mining
- Semantic subgroup discovery
- Presentation of developed tool: g-SEGS

Part II Learning from description logics (presenters: Agnieszka Lawrynowicz, Jedrzej Potoniec) Download slides RMonto website

- Refinement operators for DL-learning
- Concept learning
- Frequent pattern mining in DLs
- Similarity-based learning (e.g. kernel methods, clustering)
- An overview of example tasks (e.g. ontology evolution, semantic query results aggregation)
- Presentation of developed tool for DL-learning

Part III Semantic meta-mining (presenters: Melanie Hilario, Alexandros Kalousis) Download slides

- Meta-mining problem definition
- Goals and applications of data mining ontologies
- Representing data mining tasks, algorithms, and workflows in a DM ontology
- Ontology-based pattern extraction from data mining workflows
- Kernel based approaches
- Combining dataset, algorithm and workflow descriptors in meta-mining

Target audience

The target audience of the tutorial includes:

- Researchers in machine learning and data mining with interest in the Semantic Web technologies/ontologies
- Researchers interested in meta-mining
- Researchers interested in relational data mining/inductive logic programming
- Developers of data mining applications that would like to exploit Semantic Web technologies/ontologies to solve data mining and/or machine learning tasks

The tutorial does not require additional prior knowledge from average ECML PKDD 2011 participant.

Information about the presenters

Prof. Nada Lavrac is Head of the Department of Knowledge Technologies (since 2004), was Head of Intelligent Data Analysis and Computational Linguistics research group (in 1999- 2003) at the Department of Intelligent Systems, and researcher of Jožef Stefan Institute, Ljubljana, Slovenia (since 1978). She is Full Professor at University of Nova Gorica and Deputy Head of Information and Communication Technologies Program at Jozef Stefan International Postgraduate School. She was visiting professor at Bristol University, UK (1997-2002, teaching parts of courses Introduction to

Machine Learning and Learning from Structured Data) and at Klagenfurt University, Austria (1987-2002, teaching courses on Knowledge Acquisition, Data Mining and Decision Support). In 1984 she was in a group of researchers who were awarded a national prize for research excellence, in 1997 she was awarded the Ambassador of Science of Slovenia prize, and in 2007 she has been elected ECCAI Fellow. Her main research interest is machine learning and data mining, in particular inductive logic programming and intelligent data analysis in medicine. She is coauthor of KARDIO: A Study in Deep and Qualitative Knowledge for Expert Systems, The MIT Press 1989, and Inductive Logic Programming: Techniques and Applications, Ellis Horwood 1994, and co-editor of Relational Data Mining, Springer 2001, Intelligent Data Analysis in Medicine and Pharmacology, Kluwer 1997. She was founder the Solomon European Network and acted as co-coordinator of the EU 5th Framework project Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprise (Sol-Eu-Net, IST-1999-11495, 2000-2003). She was coordinator of the European Scientific Network in Inductive Logic Programming ILPNET (1993-1996). She is member of editorial boards of Artificial Intelligence in Medicine AI Communications New Generation Computing Applied AI Machine Learning Journal and Data Mining and Knowledge Discovery. She was vicepresident of ECCAI (1996-98), and is member of the International Machine Learning Society board (IMLS, since 2001), and Artificial Intelligence in Medicine board (AIME, since 1999).

Anže Vavpetič received his BSc in computer science at the Faculty of Computer and Information Science, University of Ljubljana in 2011. Currently he is a PhD student working at the Jožef Stefan Institute at the Department of Knowledge Technologies in Slovenia. He is interested in various topics of data mining and machine learning like relational data mining, inductive logic programming and subgroup discovery.

Agnieszka Ławrynowicz is Assistant Professor at the Institute of Computing Science at Poznan University of Technology where she also did her Ph.D. on the topic of the tutorial (with distinction). Before joining academia, she worked in industry (Empolis, Bridgestone). She also holds French-Polish DESS Certificate of Ability to Manage Companies (Poznan University of Economics & University of Rennes 1). She had an EU Marie-Curie fellowship within the PERSONET project on personalization, and Web mining at the University of Ulster. Her research interests include data mining involving Semantic Web languages, and ontology engineering. She has nearly 10 years of experience in academic teaching including the subjects of computational logics, logic programming, artificial intelligence, Web technologies, business process modeling, and recently co-authored lectures on Semantic technologies and Social Networking. She has initiated and co-organized a series of international workshops on Inductive Reasoning and Machine Learning from the Semantic Web (IRMLeS) co-located with the major European Semantic Web conference (ESWC'2009-2011), and is a co-chair of ESWC'2011 track on Inductive and Probabilistic Approaches for the Semantic Web.

Melanie Hilario holds a Ph.D. in computer science from the University of Paris VI and currently works at the University of Geneva's Artificial Intelligence Laboratory. She has initiated several European research projects on neuro-symbolic integration, meta-learning, and biological text mining. She is the scientific coordinator of the ongoing EU project e-LICO, whose goal is to build a virtual data mining lab around a planner-based discovery assistant that self-improves through ontology-based meta-mining. She has served on the program committees of conferences and workshops in machine learning and data mining. She is currently an Associate Editor of the International Journal on Artificial Intelligence Tools and a member of the Editorial Board of the Intelligent Data Analysis journal.

Alexandros Kalousis did his PhD thesis on meta-learning for classification algorithm selection at the University of Geneva, where he continues as a senior researcher. He has published widely in the area of machine learning and knowledge discovery, in particular on meta-learning, data preprocessing, feature extraction, model selection and evaluation, and metric and kernel learning for complex structures. Currently his research interests include mining over learned models, meta-mining, and

metric and kernel learning. He has served on the program committees of different data mining and machine learning conference, such as ICML, KDD, ECML/PKDD, and participated on a number of European and Swiss research projects.

Jędrzej Potoniec is currently working on his MSc thesis in Institute of Computer Science at Poznań University of Technology. His interests covers various artificial intelligence problems, especially related to semantic web. He actually works on frequent pattern discovery problem and developing new similarity measures for description logic.

Selected references

[dFGLS2010] Claudia d'Amato, Nicola Fanizzi, Marko Grobelnik, Agnieszka Lawrynowicz, Vojtech Svatek (eds.). Proc. of the 2nd ESWC Workshop on Inductive Reasoning and Machine Learning for the Semantic Web (IRMLeS 2010). Heraklion, Greece, Vol 611 of of CEUR Workshop Proceeding, CEUR-WS.org, 2010.

[dFGLS2009] d'Amato, C., Fanizzi, N., Grobelnik, M., Lawrynowicz, A., and Svátek, V. (editors) Proceedings of the 1st ESWC International Workshop on Inductive Reasoning and Machine Learning on the Semantic Web (IRMLes 2009), Heraklion, Greece, June 1, 2009. Volume 474 of CEUR Workshop Proceeding, CEUR-WS.org, 2009.

[BNC00a] L. Badea and S.-H. Nienhuys-Cheng. A re_nement operator for description logics. In J. Cussens and A. Frisch, editors, Proceedings of the 10th International Conference on Inductive Logic Programming, volume 1866 of LNAI, pages 40-59. Springer, 2000.

[BNC00b] Liviu Badea and Shan-Hwei Nienhuys-Cheng. A refinement operator for description logics. In James Cussens and Alan M. Frisch, editors, ILP, volume 1866 of Lecture Notes in Computer Science, pages 40-59. Springer, 2000.

[BS07] S. Bloehdorn and Y. Sure. Kernel methods for mining instance data in ontologies. In K. Aberer and et al., editors, Proceedings of the 6th International Semantic Web Conference, ISWC2007, volume 4825 of LNCS, pages 58-71. Springer, 2007.

[BWH05] A. Borgida, T.J. Walsh, and H. Hirsh. Towards measuring similarity in description logics. In I. Horrocks, U. Sattler, and F.Wolter, editors, Working Notes of the International Description Logics Workshop, volume 147 of CEUR Workshop Proceedings. CEUR, Edinburgh, UK, 2005.

[CH94] W.W. Cohen and H. Hirsh. Learning the CLASSIC description logic. In P. Torasso, J. Doyle, and E. Sandewall, editors, Proceedings of the 4th International Conference on the Principles of Knowledge Representation and Reasoning, pages 121-133. Morgan Kaufmann, 1994.

[dFE06] C. d'Amato, N. Fanizzi, and F. Esposito. A dissimilarity measure for ALC concept descriptions. In Proceedings of the 21st Annual ACM Symposium of Applied Computing, SAC2006, volume 2, pages 1695-1699, Dijon, France, 2006. ACM.

[dFE08] C. d'Amato, N. Fanizzi, and F. Esposito. Query answering and ontology population: An inductive approach. In S. Bechhofer and et al., editors, Proceedings of the 5th European Semantic Web Conference, ESWC2008, volume 5021 of LNCS, pages 288-302. Springer, 2008.

[dSF08] C. d'Amato, S.; Staab, and N. Fanizzi. On the influence of description logics ontologies on conceptual similarity. In A. Gangemi and J. Euzenat, editors, Proceedings of the 16th EKAW

Conference, EKAW2008, volume 5268 of LNAI, pages 48-63. Springer, 2008.

[EFI+04] F. Esposito, N. Fanizzi, L. Iannone, I. Palmisano, and G. Semeraro. Knowledge-intensive induction of terminologies from metadata. In F. van Harmelen, S. McIlraith, and D. Plexousakis, editors, ISWC2004, Proceedings of the 3rd International Semantic Web Conference, volume 3298 of LNCS, pages 441-455. Springer, 2004.

[FdE08a] N. Fanizzi, C. d'Amato, and F. Esposito. Conceptual clustering for concept drift and novelty detection. In S. Bechhofer and et al., editors, Proceedings of the 5th European Semantic Web Conference, ESWC2008, volume 5021 of LNCS, pages 318-332. Springer, 2008.

[FdE08b] N. Fanizzi, C. d'Amato, and F. Esposito. DL-Foil: Concept learning in Description Logics. In F. Zelezn_y and N. Lavra_c, editors, Proceedings of the 18th International Conference on Inductive Logic Programming, ILP2008, volume 5194 of LNAI, pages 107-121. Springer, Prague, Czech Rep., 2008.

[FdE10a] Nicola Fanizzi, Claudia d'Amato, and Floriana Esposito. Induction of concepts in web ontologies through terminological decision trees. In Jose L. Balcazar, Francesco Bonchi, Aristides Gionis, and Michele Sebag, editors, ECML/PKDD (1), volume 6321 of Lecture Notes in Computer Science, pages 442-457. Springer, 2010.

[FdE10b] Nicola Fanizzi, Claudia d'Amato, and Floriana Esposito. Learning to rank individuals in description logics using kernel perceptrons. In Pascal Hitzler and Thomas Lukasiewicz, editors, RR, volume 6333 of Lecture Notes in Computer Science, pages 173-181. Springer, 2010.

[FIPS04] N. Fanizzi, L. Iannone, I. Palmisano, and G. Semeraro. Concept formation in expressive description logics. In J.-F. Boulicaut and et al., editors, Proceedings of the 15th European Conference on Machine Learning, ECML2004, volume 3201 of LNAI, pages 99-113. Springer, 2004.

[GKM04] P. Gottgtroy, N. Kasabov, and S. MacDonell. An ontology driven approach for knowledge discovery in biomedicine. In Proceedings of the VIII Pacific Rim International Conferences on Artificial Intelligence (PRICAI), 2004.

[HKNW09] Hilario, M., Kalousis, A., Nguyen, P., Woznica, A.: A data mining ontology for algorithm selection and meta-mining. Proceedings of the ECML/PKDD09 Workshop on 3rd generation Data Mining (SoKD-09) pp. 76–87 (2009)

[IPF07] Luigi Iannone, Ignazio Palmisano, and Nicola Fanizzi. An algorithm based on counterfactuals for concept learning in the Semantic Web. Appl. Intell., 26(2):139-159, 2007.

[JLL05] J. Jozefowska, A. Lawrynowicz, and T. Lukaszewski. Towards discovery of frequent patterns in description logics with rules. In Proc. of International Conference on Rules and Rule Markup Languages for the Semantic Web (RuleML 2005), volume 3791 of LNCS, pages 84-97. Springer, 2005.

[JLL06] J. Jozefowska, A. Lawrynowicz, and T. Lukaszewski. Frequent pattern discovery in OWL DLP knowledge bases. In Managing Knowledge in a World of Networks, Proc. of EKAW 2006, volume 4248 of LNAI, pages 287-302. Springer, 2006.

[JLL08] J. Jozefowska, A. Lawrynowicz, and T. Lukaszewski. On reducing redundancy in mining relational association rules from the Semantic Web. In Proc. of the Second International Conference on Web Reasoning and Rule Systems (RR'2008), volume 5341 of LNCS, pages 205{213. Springer, 2008.

[JLL10] J. Jozefowska, A. Lawrynowicz, and T. Lukaszewski. The role of semantics in mining frequent

patterns from knowledge bases in description logics with rules. Theory and Practice of Logic Programming, 10(3):251-289, 2010.

[Kie02] J.-U. Kietz. Learnability of description logic programs. In S. Matwin and C. Sammut, editors, Proceedings of the 12th International Conference on Inductive Logic Programming, volume 2583 of LNAI, pages 117-132, Sydney, 2002. Springer.

[KM94] J.-U. Kietz and K. Morik. A polynomial approach to the constructive induction of structural knowledge. Machine Learning, 14(2):193-218, 1994.

[La09] Agnieszka Lawrynowicz. Grouping results of queries to ontological knowledge bases by conceptual clustering. In Ngoc Thanh Nguyen, Ryszard Kowalczyk, and Shyi-Ming Chen, editors, ICCCI, volume 5796 of Lecture Notes in Computer Science, pages 504-515. Springer, 2009.

[La10] Agnieszka Lawrynowicz. Foundations of frequent concept mining with formal ontologies. In Proc. of the ECML/PKDD'10 Workshop on Third Generation Data Mining: Towards Service-oriented Knowledge Discovery (SoKD-10), pages 45-50, 2010.

[LE06a] Francesca A. Lisi and Floriana Esposito. On the missing link between frequent pattern discovery and concept formation. In Stephen Muggleton, Ram_on P. Otero, and Alireza Tamaddoni-Nezhad, editors, ILP, volume 4455 of Lecture Notes in Computer Science, pages 305-319. Springer, 2006.

[LE06b] Francesca A. Lisi and Floriana Esposito. Two orthogonal biases for choosing the intensions of emerging concepts in ontology refinement. In Gerhard Brewka, Silvia Coradeschi, Anna Perini, and Paolo Traverso, editors, ECAI, volume 141 of Frontiers in Arti_cial Intelligence and Applications, pages 765-766. IOS Press, 2006.

[Leh07a] J. Lehmann. Hybrid learning of ontology classes. In P. Perner, editor, Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition, MLDM2007, volume 4571 of LNCS, pages 883-898, Leipzig, Germany, 2007. Springer.

[Leh07b] J. Lehmann. Hybrid learning of ontology classes. In P. Perner, editor, Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition, MLDM2007, volume 4571 of LNCS, pages 883-898, Leipzig, Germany, 2007. Springer.

[Leh09] Jens Lehmann. DL-learner: Learning concepts in description logics. Journal of Machine Learning Research (JMLR), 10:2639-2642, 2009.

[LH08a] J. Lehmann and P. Hitzler. Foundations of refinement operators for description logics. In H. Blockeel and et al., editors, Proceedings of the 17th International Conference on Inductive Logic Programming, ILP2007, volume 4894 of LNCS, pages 161-174. Springer, 2008.

[LH08b] J. Lehmann and P. Hitzler. A refinement operator based learning algorithm for the ALC description logic. In H. Blockeel and et al., editors, Proceedings of the 17th International Conference on Inductive Logic Programming, ILP2007, volume 4894 of LNCS, pages 147-160. Springer, 2008.

[LH10] Jens Lehmann and Christoph Haase. Ideal downward refinement in the EL description logic. In Proceedings of the 19th international conference on Inductive logic programming, ILP'09, pages 73-87, Berlin, Heidelberg, 2010. Springer-Verlag.

[Lis08] Francesca A. Lisi. Building rules on top of ontologies for the semantic web with inductive logic programming. TPLP, 8(3):271-300, 2008.

[LKNM+09] N. Lavrac, P. Kralj Novak, I. Mozetic, V. Podpecan, H. Motaln, M. Petek, and K. Gruden. Semantic subgroup discovery: Using ontologies in microarray data analysis. In Proc. 31st Annual Intl. Conf. of the IEEE EMBS, pages 5613-5616, 2009.

[LM04] F.A. Lisi and D. Malerba. Inducing multi-level association rules from multiple relations. Machine Learning Journal, 55(2):175-210, 2004.

[NVTL09] Petra Kralj Novak, Anze Vavpetic, Igor Trajkovski, and Nada Lavrac. Towards semantic data mining with g-segs. In Proceedings of the 11th International Multiconference Information Society IS 2009, 2009.

[PJvL09] V. Podpecan, M. Jursic, M. Zakova, and N. Lavrac. Towards a service oriented knowledge discovery platform. In V. Podpecan and N. Lavrac, editors, Third-generation data mining: towards service-oriented knowledge discovery, pages 25-36, 2009.

[PSD09] Panov, P., Soldatova, L., Dzeroski, S.: Towards an ontology of data mining investigations. Lecture Notes in Artificial Intelligence 5808, 257–271 (Jan 2009)

[RNT09] Achim Rettinger, Matthias Nickles, and Volker Tresp. Statistical relational learning with formal ontologies. In Wray L. Buntine, Marko Grobelnik, Dunja Mladenic, and John Shawe-Taylor, editors, ECML/PKDD (2), volume 5782 of Lecture Notes in Computer Science, pages 286-301. Springer, 2009.

[RV00] C. Rouveirol and V. Ventos. Towards learning in CARIN-ALN. In J. Cussens and A. Frisch, editors, Proceedings of the 10th International Conference on Inductive Logic Programming, volume 1866 of LNAI, pages 191-208. Springer, 2000.

[SRR05] V. Svatek, J. Rauch, and M. Ralbovsky. Ontology-enhanced association mining. In Semantics, Web and Mining, Joint International Workshops, EWMF 2005 and KDO 2005, pages 163-179, 2005.

[TLT08] I. Trajkovski, N. Lavrac, and J. Tolar. SEGS: Search for enriched gene sets in microarray data. Journal of Biomedical Informatics, 41(4):588-601,2008.

[VS10] Joaquin Vanschoren and Larisa Soldatova. Exposé: An ontology for data mining experiments. In International Workshop on Third Generation Data Mining: Towards Service-oriented Knowledge Discovery (SoKD-2010), September 2010.