Propositionalization Online

Nada Lavrač[1,2,3], Matic Perovšek[1,2], and Anže Vavpetič[1,2]

[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[3] University of Nova Gorica, Nova Gorica, Slovenia
{nada.lavrac,matic.perovsek,anze.vavpetic}@ijs.si

Abstract. Inductive Logic Programming and Relational Data Mining address the task of inducing models or patterns from multi-relational data. An established relational data mining approach is propositionalization, characterized by transforming a relational database into a single-table representation. The paper presents a propositionalization toolkit implemented in the web-based data mining platform ClowdFlows. As a contemporary integration platform it enables workflow construction and execution, provides open access to Aleph, RSD, RelF and Wordification feature construction engines, and enables RDM performance comparison through cross-validation and ViperCharts results visualization.

Keywords: relational data mining, propositionalization, web access.

1 Introduction

Propositional data mining algorithms induce hypotheses in the form of models or patterns learned from a given data table. In contrast, Inductive Logic Programming (ILP) [6] and Relational Data Mining (RDM) [1] algorithms induce models or patterns from multi-relational data (e.g., relational databases). For relational databases with clearly identifiable instances (i.e., *individual-centered representations* [2], characterized by one-to-many relationships among data tables), propositionalization techniques [3] can be used to transform a relational database into a propositional single-table format, followed by propositional learning, e.g., by using a decision tree or a classification rule learner.

This paper presents an online propositionalization toolkit, which can be used to construct RDM workflows. As completed workflows, data, and results can be made public by the author of the workflow, the platform can serve as an easy-to-access integration platform for various RDM workflows.

2 Clowdflows ILP module

The ClowdFlows platform [4] is an open-source, web-based data mining platform that supports the construction and execution of scientific workflows. This web application can be accessed and controlled from anywhere while the processing is performed in a cloud of computing nodes. A public installation of ClowdFlows

T. Calders et al. (Eds.): ECML PKDD 2014, Part III, LNCS 8726, pp. 456–459, 2014.

is accessible at http://clowdflows.org. For a developer, the graphical user interface supports simple operations that enable workflow construction: adding workflow components (widgets) on a canvas and creating connections between the components to form an executable workflow, which can be shared by other users or developers. Upon registration, the user can access, execute, modify, and store the modified workflows, enabling their sharing and reuse. On the other hand, by using anonymous login, the user can execute a predefined workflow, while any workflow modifications would be lost upon logout.

We have extended ClowdFlows with the implementation of an ILP toolkit, including the popular ILP system Aleph [9] together with its feature construction component, as well as RSD [10], RelF [5] and Wordification [7] propositionalization engines. Construction of RDM workflows is supported by other specialized RDM components (e.g., the MySQL package providing access to a relational database by connecting to a MySQL database server), other data mining components (e.g., the Weka classifiers) and other supporting components (including cross-validation), accessible from other ClowdFlows modules. Each public workflow is assigned a unique URL that can be accessed by any user to either repeat the experiment, or use the workflow as a template to design another workflow. Consequently, the incorporated RDM algorithms become handy to use in real-life data analytics, which may therefore contribute to improved accessibility and popularity of ILP and RDM.

Figure 1 shows two simple workflows using the ILP and Weka module components. The first workflow assumes that the user uploads the files required by RSD

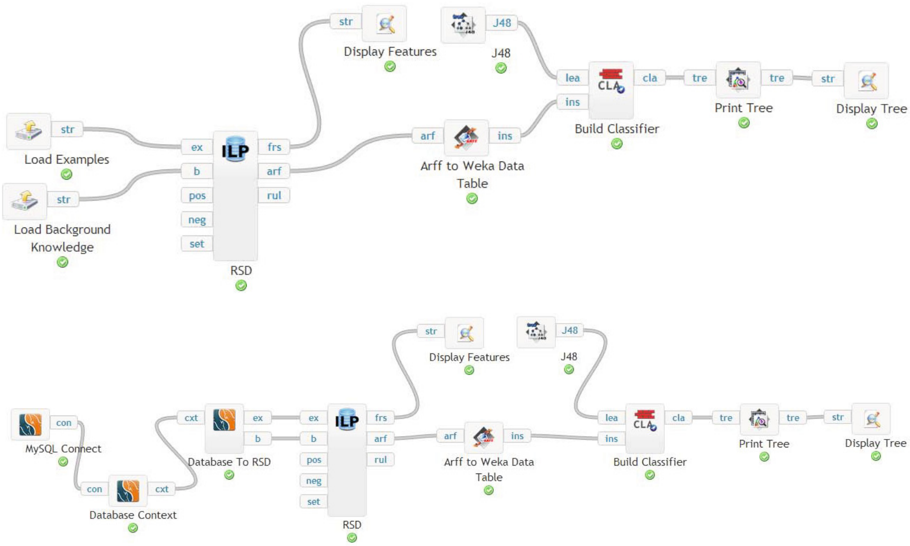

Fig. 1. Above: Simple RSD propositionalization workflow using ILP and Weka components, available online at http://clowdflows.org/workflow/471/. Below: The same RSD workflow, extended by accessing the training data using a MySQL database, available at http://clowdflows.org/workflow/611/.

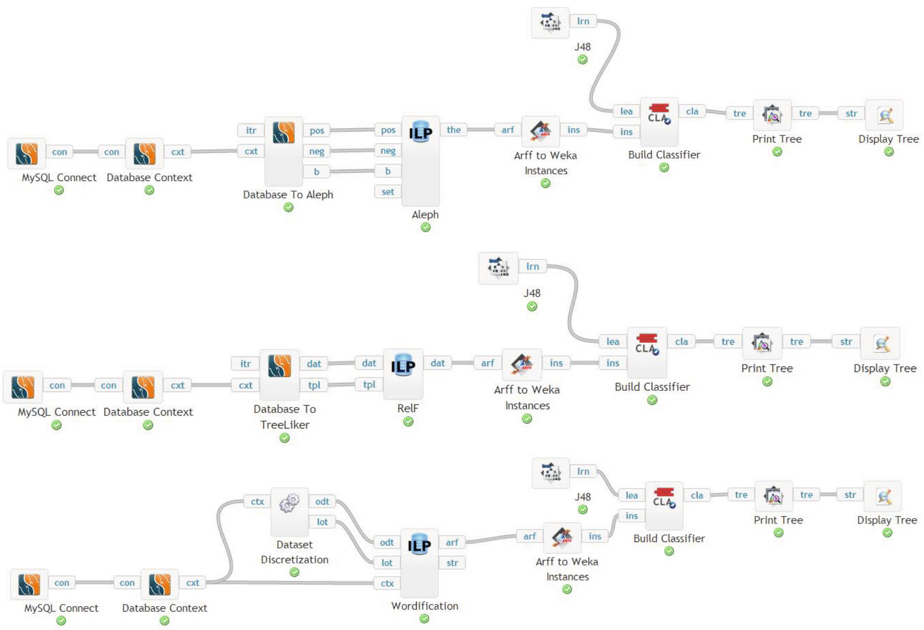

Fig. 2. Propositionalization workflows available online: for Aleph at `http://clowdflows.org/workflow/2224/`, for RelF at `http://clowdflows.org/workflow/2227/` and for Wordification at `http://clowdflows.org/workflow/2222/`.

as Prolog programs, while the second workflow extends this use case by retrieving the training data from a MySQL database server and automatically constructing the background knowledge and the training examples. Similar workflows, constructed for the other three propositionalization approaches Aleph, RelF and Wordification, are illustrated in Figure 2.

The evaluation workflow is shown in Figure 3. After reading the relational data and data discretization, propositionalization algorithms are applied, their results are transformed into the Weka input format for the J48 decision tree learner, followed by 10-fold cross-validation with identical folds allowing performance comparison of different propositionalization algorithms. The results of cross-validation (precision, recall, F-score) are connected to the input of VIPER (Visual Performance Evaluation) engine [8], which displays the results as points in the precision-recall space. The evaluation workflow enables ILP researchers to reuse the developed workflow and its components in future experimentation.

In terms of workflows reusability, accessible by a single click on a web page where a workflow is exposed, the implemented propositionalization toolkit is a significant step towards making the ILP legacy accessible to the research community in a systematic and user-friendly way. To the best of our knowledge, this is the only workflow-based implementation of ILP and RDM algorithms in a platform accessible through a web browser, enabling simple workflow adaptation to the user's needs.

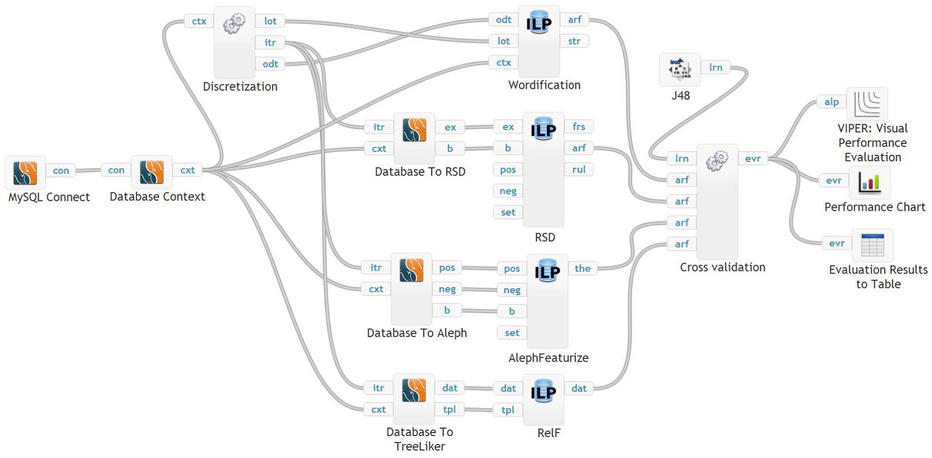

Fig. 3. Performance evaluation workflow, available at `http://clowdflows.org/workflow/2210/`, comparing the results of J48 after propositionalization by Aleph, RSD, RelF and Wordification.

References

[1] Džeroski, S., Lavrač, N. (eds.): Relational Data Mining. Springer (2001)
[2] Flach, P.A., Lachiche, N.: 1BC: A First-Order Bayesian Classifier. In: Džeroski, S., Flach, P.A. (eds.) ILP 1999. LNCS (LNAI), vol. 1634, pp. 92–103. Springer, Heidelberg (1999)
[3] Kramer, S., Lavrač, N., Flach, P.A.: Propositionalization approaches to relational data mining. In: Džeroski and Lavrač pp. 262–292
[4] Kranjc, J., Podpečan, V., Lavrač, N.: ClowdFlows: A cloud based scientific workflow platform. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2012, Part II. LNCS, vol. 7524, pp. 816–819. Springer, Heidelberg (2012)
[5] Kuželka, O., Železný, F.: Block-wise construction of tree-like relational features with monotone reducibility and redundancy. Machine Learning 83(2), 163–192 (2011)
[6] Muggleton, S. (ed.): Inductive Logic Programming. Academic Press, London (1992)
[7] Perovšek, M., Vavpetič, A., Cestnik, B., Lavrač, N.: A wordification approach to relational data mining. In: Fürnkranz, J., Hüllermeier, E., Higuchi, T. (eds.) DS 2013. LNCS, vol. 8140, pp. 141–154. Springer, Heidelberg (2013)
[8] Sluban, B., Lavrač, N.: ViperCharts: Visual performance evaluation platform. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds.) ECML PKDD 2013, Part III. LNCS, vol. 8190, pp. 650–653. Springer, Heidelberg (2013)
[9] Srinivasan, A.: Aleph manual (March 2007), `http://www.cs.ox.ac.uk/activities/machinelearning/Aleph/`
[10] Železný, F., Lavrač, N.: Propositionalization-based relational subgroup discovery with RSD. Machine Learning 62(1-2), 33–63 (2006)