

SOFIA: An Analytics Recommendation System

Fatemeh Nargesian, Alain Biem, Prateek Jain, Srinivasan Parthasarathy, and
Deepak S. Turaga

Department of Computer Science, University of Toronto, IBM Watson Research Center
fnargesian@cs.toronto.edu, abiem@yahoo.com, jainprateek@gmail.com,
com, {spartha, turaga}@us.ibm.com

Abstract. The SOFIA analytic recommender system helps non-expert users to select analytics (algorithms and their implementations) to fulfill a data mining task in an effective manner. The recommender is a novel framework that applies an ontology-driven approach to recommend a ranked list of analytics to fulfill a task based on the characteristics of the given dataset. SOFIA relies on an ontological representation of data science principles evolved by the data science community; it does not require training examples nor actual deployment of candidate analytics on given datasets.

Keywords: Analytics Recommendation; Data mining; Ontology; Meta-learning

1 Introduction

In several domains, such as data mining, there is a variety of algorithms and their implementations, in major software packages, that can be considered as candidates to solve particular data mining problems. One of the most difficult tasks in such domains is to decide when one analytic (an algorithm or one of its implementations) is better than the other to do a data mining task on a dataset with particular characteristics. “Meta-learning” is the automatic process of knowledge acquisition that relates the performance of the learning algorithms with the characteristics of the machine learning problems and datasets [4]. SOFIA is an application which delivers the following services: registering data mining analytics (algorithms or executables) by developers and experts, querying and searching analytics (keyword search), visualizing analytics, deploying single executables or composed analytics, and recommending analytics. In this paper, we introduce the analytics recommendation aspect of SOFIA, demonstrated in Figure 2. The users of SOFIA are data analysts in different domains who are interested in finding the most efficient and effective analytics to do data mining tasks. These users have limited knowledge about data mining algorithms, implementations and the characteristics of their datasets.

Several approaches to solving the problem of analytics selection can be identified. The most straightforward approach is to evaluate each available analytic, on the given dataset, and select the one yielding the best results for the task. This data-driven approach is simple and intuitive but is time-consuming when the data set is large and there is a large choice of analytics for the task [2]. A more efficient way is the use of expert or consensus knowledge relating the characteristics of analytics, data mining tasks and datasets [1]. The use of this knowledge is expressed via reasoning rules; however, developing rules for analytics selection may be expensive and unfeasible, since good human

experts are not always available [2]. SOFIA has unified the consensus knowledge about different data mining algorithms and analytics in order to support automatic knowledge discovery and provide a base for research. SOFIA ontology is an OWL-based ontology which represents machine learning analytics at various levels of abstractions in a taxonomy. At the high level, the SOFIA ontology represents general data mining tasks such as classification, clustering, or prediction, etc. Each task can then be achieved by a set of algorithms and each algorithm has various implementations registered in the system. The SOFIA ontology, currently, consists of 342 implementations in languages such as Python, Java and C++ for 461 algorithms and 227 number of data mining tasks. SOFIA automatically derives the characteristics of a dataset by analyzing its schema and data values and infers more characteristics following some ontological rules embedded in the ontology. Then, a set of recommendation rules are generated on the fly to match the characteristics of SOFIA analytics and those of the given dataset to come up with recommendations. In SOFIA, the logic of analytics and dataset property matching is expressed, by a data mining expert, as simple analytic-to-dataset characteristics mappings. A mapping is an association between a dataset's property and its value and an analytic's property and its value. It describes a partial condition for matching analytics to datasets. An instance of an analytic-to-dataset mapping is: if an algorithm supports multi-category classification, it could be deemed as a recommendation for a dataset with more than two class labels. A composition of mappings between dataset properties and analytic properties discovers whether the analytic would perform a specific task efficiently and effectively on the dataset. An example of such composition is shown in Figure 1(b) as a rule in Semantic Web Rule Language (SWRL)(<http://protege.stanford.edu>). According to this rule, it is inferred that, in Figure 1(a), a classification algorithm A with shown characteristics is a good match for dataset D , since the algorithm A is relatively robust to noise and missing data, thus, it can efficiently handle the large amount of noise and missing data in dataset D . Moreover, the input and class type of algorithm A is similar to those of dataset D . In SOFIA, other properties of analytics, such as efficiency, numerical stability, etc. are also taken into account while matching a dataset to analytics. SOFIA adopts an algorithm for automatic compilation of mappings and generating recommendation rules. These rules are created dynamically on the fly according to the dataset and data mining task provided by the user. After incorporating the rules in SOFIA ontology, the reasoning algorithm of SOFIA evaluates the rules against the ontology and the recommendations are inferred.

2 SOFIA Analytics Recommender

Sofia analytics recommender has two modes for different users: (1) domain experts, (2) data analysts. The domain expert mode allows data mining experts of SOFIA to modify dataset-analytics mappings and semantic general rules in text format. This application mode of SOFIA is not demonstrated in this paper due to lack of space. In data analyst mode, the user first specifies a dataset by uploading a dataset in CSV format or selecting from a list of existing datasets (UCI repository). Then, she selects a data mining task and the number of required recommendations as shown in left part of Figure 2. Based on the selected task, the system might ask the user for extra information. For instance, in a classification task, SOFIA needs to know which column in the CSV file is the class

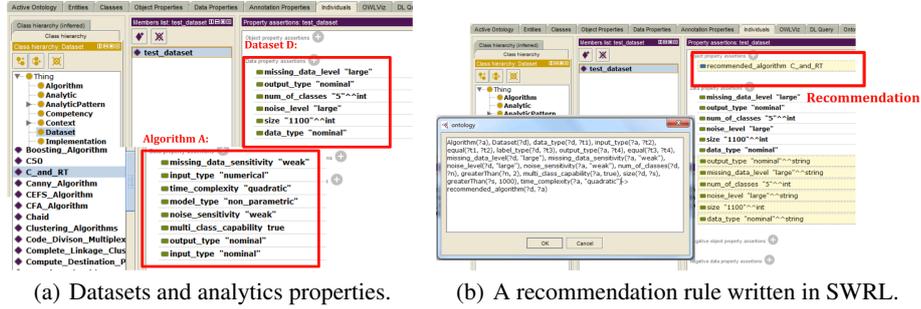


Fig. 1: Matching datasets and analytics.

label column. SOFIA extracts native, derived and task-specific properties of the dataset such as size, missing data level, noisy data level, etc. These properties are shown to the user in case a user would like to do further corrections on dataset properties, as in the middle part of Figure 2. In the next step, the recommender composes a set of the recommendation rules, on the fly, based on the characteristics of the dataset and selected task. Then, the recommendation rules are incorporated into SOFIA ontology. The recommender uses Pellet reasoner (<http://clarkparsia.com/pellet>) to evaluate the rules. These analytics are then ranked by the sum of the mappings' significance scores involving in the rule that infers them [3]. In the following sections, we describe the dataset properties extraction and rule composition components of SOFIA recommender in more details.

2.1 Dataset Properties Extraction

SOFIA represents a dataset with a set of native (e.g. data type, size), derived (e.g. missing data level, noisy data level) and task specific properties (e.g. number of classes, class label type for classification task). Dataset properties extraction module scans the dataset provided by the user and generates the dataset profile as data property values. For example, missing data ratio is the average number of missing data values in each row of a dataset. This ratio can then be translated into a string value for *missing data level* property (e.g., *small*, *medium* and *large*) assuming threshold intervals.

2.2 Automatic Generation of Recommendation Rules

Manually creating recommendation rules that result in high quality recommendations costs domain experts' effort and time. SOFIA takes as input a set of simple mappings, written by an expert, and generates a set of recommendation ontological rules, automatically and on the fly, by combining appropriate mappings. The body of a recommendation rule is a conjunction of conditions on dataset and analytics' properties that qualify analytics as recommendations for a dataset. The rule generation module of SOFIA adopts a filtering approach to prune candidate analytics for recommendation. The algorithm starts by assuming that all analytics available in the ontology are potential recommendations. The mapping selection procedure follows a greedy policy to select mappings based on significance scores. These scores are provided in the mapping file by the domain expert. At each step, a mapping is selected and if its premise is satisfied

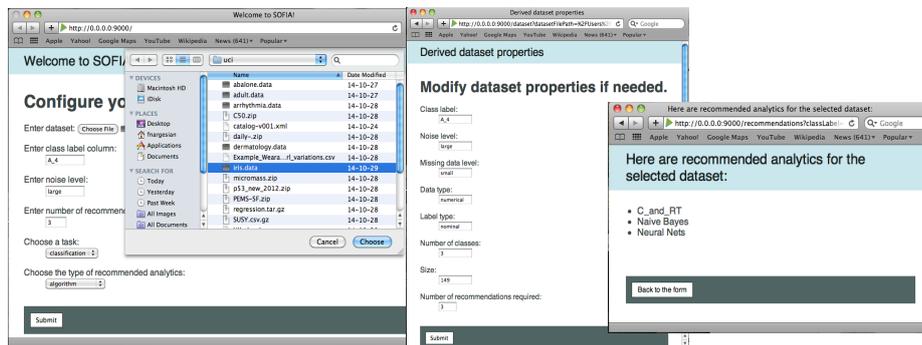


Fig. 2: The data analyst inputs a dataset and a data mining task (left) the derived dataset properties are confirmed by the user (middle) the recommended analytics are returned to the user.

based on the characteristics of the dataset, the clauses in its head are added to the partial recommendation rule. The head clauses are mostly related to the characteristics of recommended analytics. Then, the partial rule is added to the ontology and the reasoner infers recommended analytics that might be filtered in the next filtering steps. The filtering procedure terminates when the partial rule is restrictive enough to return the number of required recommendations.

3 Implementation and Demonstration

In this demonstration, we introduce SOFIA, a tool to explore data mining analytics and request for recommendations for fulfilling a task on a user given dataset. SOFIA has been implemented using Java. SOFIA ontology represents machine learning analytics using Web Ontology Language (OWL). SOFIA offers an interactive web interface implemented using play framework. In the demonstration, users can play with SOFIA ontology visualization tool order to become familiar with existing analytics' properties, available at different levels of taxonomy. They can also browse description and ontological information about different analytics and search for analytics using keywords. Users can also advocate expert users and register new analytics or matching mappings to SOFIA. On the other hand, a user can use SOFIA as a data analyst requesting analytics recommendations for fulfilling a data mining task on a dataset. In this demonstration, we will highlight features of SOFIA using both UCI repository datasets (<https://archive.ics.uci.edu/ml/datasets.html>) and real datasets.

References

1. Christophe Giraud-Carrier et al. Introduction to the special issue on meta-learning. In *Machine Learning*, volume 54, pages 187–193, 2004.
2. Ricardo B. C. Prudncio et al. Meta-learning approaches to selecting time series models. In *Machin Learning*, pages 121–137, 2004.
3. et al. Quan Sun. Pairwise meta-rules for better meta-learning-based algorithm ranking. In *Machine Learning*, volume 93, pages 141–161, 2013.
4. et al. Smith-Miles. Cross-disciplinary perspectives on meta-learning for algorithm selection. In *ACM Computer Survey*, volume 41, 2009.

Title of the Book or Conference Name: The Semantic Web - ISWC 2015, 14th International Semantic Web Conference, Bethlehem, Pennsylvania, United States, October 11-15, 2015

Volume Editor(s): Marcelo Arenas, Oscar Corcho, Elena Simperl, Markus Strohmaier, Mathieu d'Acquin, Kavitha Srinivas, Paul Groth, Michel Dumonier, Jeff Heflin, Krishnasasad Thirunarayan, Steffen Staab

Title of the Contribution: . . . SOFIA : . . . An Analytics Recommendation System

Author(s) Name(s): Fatemeh Nargesian, Alain Bism, Prateek Jain, Srinivasan Parthasarathy, Deepak

Corresponding Author's Name, Address, Affiliation and Email: . . . Fatemeh . . . Nargesian . . . - ^{5 Tureya}

. . . Department . of . . . Computer . Science . . . University . of . Toronto . . . - . . . fnargesian@cs.toronto.edu

. . . Apt. 814 . . . 761 . . . Bay . Street . . . Toronto . . . M5G . 2R2 . . . Canada

When Author is more than one person the expression "Author" as used in this agreement will apply collectively unless otherwise indicated.

§ 1 Rights Granted

Author hereby grants and assigns to Springer International Publishing AG, Cham (hereinafter called Springer) the exclusive, sole, permanent, world-wide, transferable, sub-licensable and unlimited right to reproduce, publish, distribute, transmit, make available or otherwise communicate to the public, translate, publicly perform, archive, store, lease or lend and sell the Contribution or parts thereof individually or together with other works in any language, in all revisions and versions (including soft cover, book club and collected editions, anthologies, advance printing, reprints or print to order, microfilm editions, audiograms and videograms), in all forms and media of expression including in electronic form (including offline and online use, push or pull technologies, use in databases and networks for display, print and storing on any and all stationary or portable end-user devices, e.g. text readers, audio, video or interactive devices, and for use in multimedia or interactive versions as well as for the display or transmission of the Contribution or parts thereof in data networks or search engines), in whole, in part or in abridged form, in each case as now known or developed in the future, including the right to grant further time-limited or permanent rights. For the purposes of use in electronic forms, Springer may adjust the Contribution to the respective form of use and include links or otherwise combine it with other works. For the avoidance of doubt, Springer has the right to permit others to use individual illustrations and may use the Contribution for advertising purposes.

The copyright of the Contribution will be held in the name of Springer. Springer may take, either in its own name or in that of copyright holder, any necessary steps to protect these rights against infringement by third parties. It will have the copyright notice inserted into all editions of the Contribution according to the provisions of the Universal Copyright Convention (UCC) and dutifully take care of all formalities in this connection in the name of the copyright holder.

§ 2 Regulations for Authors under Special Copyright Law

The parties acknowledge that there may be no basis for claim of copyright in the United States to a Contribution prepared by an officer or employee of the United States government as part of that person's official duties. If the Contribution was performed under a United States government contract, but Author is not a United States government employee, Springer grants the United States government royalty-free permission to reproduce all or part of the Contribution and to authorize others to do so for United States government purposes.

If the Contribution was prepared or published by or under the direction or control of Her Majesty (i.e., the constitutional monarch of the Commonwealth realm) or any Crown government department, the copyright in the Contribution shall, subject to any agreement with Author, belong to Her Majesty.

If the Contribution was created by an employee of the European Union or the European Atomic Energy Community (EU/Euratom) in the performance of their duties, the regulation 31/EEC, 11/EAEC (Staff Regulations) applies, and copyright in the Contribution shall, subject to the Publication Framework Agreement (EC Plug), belong to the European Union or the European Atomic Energy Community.

If Author is an officer or employee of the United States government, of the Crown, or of EU/Euratom, reference will be made to this status on the signature page.

§ 3 Rights Retained by Author

Author retains, in addition to uses permitted by law, the right to communicate the content of the Contribution to other scientists, to share the Contribution with them in manuscript form, to perform or present the Contribution or to use the content for non-commercial internal and educational purposes provided the Springer publication is mentioned as the

