An Ontological Approach for Recommending a Feature Selection Algorithm

Chapter · January 2022

DOI: 10.1007/978-3-031-09917-5_20

3 authors:

Aparna Nayak
TU Dublin
14 PUBLICATIONS **324** CITATIONS
SEE PROFILE

Bojan Bozic
TU Dublin
18 PUBLICATIONS **81** CITATIONS
SEE PROFILE

Luca Longo
TU Dublin
151 PUBLICATIONS **2,825** CITATIONS
SEE PROFILE

An Ontological Approach for Recommending a Feature Selection Algorithm

Aparna Nayak(✉), Bojan Božić, and Luca Longo

SFI Centre for Research Training in Machine Learning, School of Computer Science, Technological University Dublin, Dublin, Republic of Ireland
{aparna.nayak,bojan.bozic,luca.longo}@tudublin.ie

Abstract. Feature selection plays an important role in machine learning or data mining problems. Removing irrelevant features increases model accuracy and reduces the computational cost. However, selecting important features is not a simple task as one feature selection algorithm does not perform well on all the datasets that are of interest. This paper tries to address the recommendation of a feature selection algorithm based on dataset characteristics and quality. The research uses three types of dataset characteristics along with data quality metrics. The main contribution of the work is the utilization of Semantic Web techniques to develop a novel system that can aid in robust feature selection algorithm recommendations. The system's strength lies in assisting users of machine learning algorithms by providing more relevant feature selection algorithms for the dataset using an ontology called Feature Selection algorithm recommendation based on Data Characteristics and Quality (FSDCQ). Results are generated using six different feature selection algorithms and four types of classifiers on ten datasets from UCI repository. Recommendations take the form of "Feature selection algorithm X is recommended for dataset i, as it performed better on dataset j, similar to dataset i in terms of class overlap 0.3, label noise 0.2, completeness 0.9, conciseness 0.8 units". While the domain-specific ontology FSDCQ was created to aid in the task of algorithm recommendation for feature selection, it is easily applicable to other meta-learning scenarios.

Keywords: Feature selection algorithms · Meta-features · Ontology

1 Introduction

Feature selection is one of the core phases of any machine learning (ML) task, as it might significantly improve model building by removing irrelevant features. Several algorithms have been developed for such a phase and choosing one among the many is a costly decision, a trade-off between the time spent by automatic procedures and domain experts [4,7]. Inappropriate feature selection algorithms might cause serious problems, such as compromising the quality of the patterns

T. Di Noia et al. (Eds.): ICWE 2022, LNCS 13362, pp. 300–314, 2022.
https://doi.org/10.1007/978-3-031-09917-5_20

to be learnt from data and, thus, model performance. A common approach is 'trial-and-error', which tends to be often effective [19]. Another approach is to choose a feature selection algorithm based on the characteristics of the dataset. Specifically, this can be implemented by using meta-learning concepts [36] and by utilizing dataset characteristics that are called "meta-features". Automating the algorithm selection process for feature selection is a challenge in data mining. However, if overcome, it has the potential to significantly increase data scientists and machine learning practitioners productivity [24]. There exists a relationship between the performance of a feature selection algorithm and the characteristics of the dataset [32].

To address this specific relationship, we propose a domain ontology along with the consideration of Dataset Characteristics and Quality (DCQ), respectively representing dataset characteristics and the quality of information. Feature Selection algorithm recommendation using DCQ (FSDCQ), is modeled by adding rules to the domain ontology DCQ, to enhance the expressivity which acts as a recommender. The benefits of using an ontology to deliver such a recommendation include interoperability, potential reuse, and sharing of knowledge [35]. The particular research question investigated in this research is: "To what extent can a domain ontology facilitate the recommendation of feature selection algorithms?". The work's main objective is the adoption of Semantic Web techniques to develop a novel system that can aid in robust feature selection algorithm recommendation. The use of rule languages enables a better understanding of the role of each meta-feature, thereby increasing the model's explainability [13,39,40].

The remainder of this article is structured as follows. Section 2 reviews related work on the existing approaches to automatically recommend feature selection algorithms, and existing ontologies to describe the dataset quality and its characteristics. Section 3 presents a novel domain ontology, followed by a description of an empirical experiment in Sect. 4. Results of such an experiment are presented and discussed in Sect. 5. Finally, Sect. 6 concludes the research work by providing directions for future work.

2 Related Work

This section briefly discusses the existing work on automatic feature selection recommendation methods and the application of ontologies related to data characteristics and its quality.

2.1 Feature Selection

The two primary feature selection methods identified include (i) the filter approach and (ii) the wrapper approach. While various feature selection algorithms have been proposed, some of these outperform others in terms of performance (for example, classification accuracy) for a given dataset [41]. This results in the emergence of a new research area devoted to establishing intrinsic relationships

between dataset characteristics and feature selection algorithms. A literature review was carried out in order to identify techniques that recommend a feature selection algorithm based on meta-features. Meta-features, describe the properties of the dataset which are predictive for the performance of machine learning algorithms trained on them [29]. The description of a dataset in terms of its information/statistical properties can be referred to as dataset characteristics. Three distinct sets of measures are used to extract dataset characteristics: (i) simple, statistical, and information-theoretic features (ii) model-based features (iii) landmarking features [38]. Simple properties represent those taken from the attribute value table of the dataset. Statistical properties are used to determine the correlation and symmetry of attributes. Information-theoretical properties seek to characterise the nominal attributes and their relationship with the class attribute. Model-based properties adopt ML methods to represent datasets. Landmarking properties illustrate the performance achieved by simple classification algorithms.

Table 1 summarises the literature covering those approaches in which meta-features were used to build a recommendation model for automatically selecting algorithms in machine learning. In detail, an advisory function refers to a method that aims to recommend an algorithm from an existing knowledge base. The proposed work aims to use ontology as advisory function. Some of the applications that uses ontology as advisory methods/recommendation are, product recommendation based on text [31], health-care [5,6], higher education [17]. Therefore, it is a novel approach to solve recommendation of feature selection algorithm using ontology. To the best of our knowledge, no research has focused on considering data quality as a characteristic of a dataset. In this article, beside the aforementioned simple, statistical, information, and quality-based measures we propose an additional category to characterise datasets, which includes quality-based measures.

2.2 Ontology

A methodology to build an ontology from scratch is discussed in Methontology [8] where a set of activities conforming the ontology development process is presented. Following best practices in ontology development, the Data Characteristics and Quality (DCQ) ontology reuses appropriate classes from a set of ontologies that are designed for data quality and data mining applications. An extensive literature has been conducted to understand existing vocabularies to support meta-features, and a vocabulary of terms have been composed for DCQ.

Meta-features are usually described as a part of Data Mining (DM) ontologies. 'OntoDM' is a general data mining ontology designed to provide a unified framework for data mining research. It makes an attempt to encompass the entirety of the data mining cycle [20]. 'Expose' is an ontology for standardizing the description of machine learning experiments. This ontology is used to express and share meta-data about experiments [37].

Table 1. Literature review and comparison of advisory functions used for recommendations

Source	Advisory function	Number of datasets	Number of classification techniques	Number of feature selection algorithms	Evaluation metrics	Dataset characteristic			
						Simple, statistical	Information theoretical	Model based	Land marking
[11]	Ranking based on McNemar test	1082*	5	8	Accuracy	✓	✓	✗	✗
[14]	SVM	156	–	7	Accuracy	✓	✓	✗	✗
[15]	kNN	58	–	–	F1 score				
[19]	C5.0 decision tree	128	5	3	Accuracy, time complexity	✓	✓	✗	✗
[23]	Ranking based on MCPM	213	5	5	Learning time, Percentage of selected attributes, Error rate	✓	✓	✓	✓
[25]	kNN	47	–	10	Spearman's rank correlation	✓	✓	✗	✓
[26]	kNN	38	–	9	Accuracy	✓	✓	✗	✗
[27]	Regression	123	–	5	Correlation	✓	✓	✓	✗
[28]	Regression	54	–	9	Accuracy	✓	✓	✓	✓
[32]	J4.8 decision trees	26	4	3	Accuracy	✓	✗	✗	✗
[33]	kNN	84	–	–	Accuracy, Execution time	✓	✗	✗	✗
[41]	kNN	115	22	5	Recommendation hit ration based on accuracy	✓	✓	✗	✗
[43]	Variance, LIBSVM	84	–	3	Accuracy	✓	✓	✓	✓

* includes artificial dataset

To represent the relationship between data mining tasks and dataset characteristics, multiple ontologies have been designed. 'OntoDM-KDD' [21], 'OntoDT' [22], 'CRISP-DM' [34] are some of the additional ontologies that are based on data mining-related concepts. 'DMOP' is a data mining optimization ontology that supports various stages of the data mining process [12]. A class hierarchy that relates datasets and their features that were established in DMOP is reused in DCQ.

Data quality is one of the essential component while describing a dataset. Data Quality Management (DQM) is an ontology that refers to the conceptualization of the data quality domain, the establishment of cleaning standards, and the reporting of data quality problems [9]. Data Cleaning Ontology (DCO) refines and extends data cleaning operations which directly assesses data quality [2]. Reasoning Violations Ontology (RVO) describes the reasoning errors of RDF and OWL [3]. Another matured ontology is recommended by the World Wide Web Consortium (W3C)[1] which covers most of the aspects of data quality [1].

3 A Novel Ontological Model

In order to recommend feature selection algorithms intelligently by extracting meta-features from a dataset, reuse of classes from existing ontologies is proposed. Specifically, the proposed ontology is developed by considering and

[1] https://www.w3.org/TR/vocab-dqv/.

reusing classes from the 'OntoDT', 'OntoDM-KDD', 'CRISP-DM' ontologies along with the 'DCO', 'DQM', 'RVO', and 'DQV' ontologies. The W3C recommendation ontology language, OWL (Web Ontology Language), is adopted to develop such an ontology with Protégé editor.

3.1 Feature Selection Algorithm Recommendation Using Dataset Characteristics and Quality (FSDCQ) Ontology

Over the last several decades, researchers in meta-learning have actively investigated data characteristics that may aid in the development of models. The DQV ontology proposes categories, dimensions, and metrics for data quality, and a similar approach is used in DCQ, where data characteristics are viewed as metrics. These metrics are classified into five dimensions, which fall under the dataset characteristics and quality category as shown in Tables 4 and 5. The class hierarchy of the FSDCQ ontology is shown in Fig. 1. Table 2 depicts ontology metrics of FSDCQ before adding individuals.

The data characteristics and quality vocabulary requirements are specified with a set of competency questions. Competency questions also help users evaluate an ontology. To develop competency questions, we must first define our domain of interest, for which our ontology will serve as a representation. Information gathering is a critical component to accomplishing this goal, especially if we do not fully understand the subject matter for which we are developing an ontology. FSDCQ is primarily concerned with conceptualizing the relationship between meta-features and a feature selection algorithm.

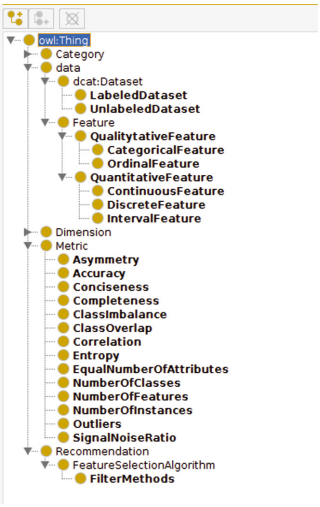


Fig. 1. Class hierarchy of FSDCQ

Table 2. FSDCQ metrics

Property	Count
Axioms	396
Classes	39
Logical axioms	326

Competency questions are directed at users and help us define the scope of an ontology. In other words, these are the queries for which users search an ontology and its associated knowledge base for solutions. The following are the main competence questions linked with proposed FSDCQ:

- **CQ:** Given a machine learning classification task/dataset, which feature selection algorithm will yield optimal results? This competency question is decomposed into many sub-questions. Coarse-grained questions include
  - **CQa:** Given only a set of pieces of data quality information, which feature selection algorithm performs the best?
  - **CQb:** Given only a set of pieces of data characteristics information, which feature selection algorithm performs the best?

The competency questions, at a more granular level, are listed in Table 3. These questions can be queried on the FSDCQ ontology using SPARQL to understand whether the modeled ontology meets the user requirements.

4 Proposed Methodology

This section presents a recommendation model for feature selection algorithm, as depicted in Fig. 2. The implementation process is divided into three main steps, as detailed below:

Table 3. Competency questions of Feature Selection algorithm recommendation using Dataset Characteristics and Quality ontology

CQ2: What characteristics belong to a dataset?
CQ3: What are the different measures to compute data quality for classification tasks?
CQ4: Which feature selection algorithm is suitable for reaching the data quality level X?
CQ5: What are the dataset characteristics that a feature selection method X requires?

- extraction of dataset characteristics and quality information;
- formation of a rule base using feature selection algorithms;
- populating ontology for the recommendations

These steps are described in details in the following sections.

4.1 Extraction of Dataset Characteristics and Quality

The dataset repository contains multiple datasets from which meta-features are extracted. Flat files are used in the experiment, which contain lines of text extracted from a collection of uniform records, each of which contains multiple attributes separated by a comma, semicolon, space, or tab.

1. Preprocessing: This is the first phase in which raw dataset is considered as input. Headers in the original dataset are not considered for analysis. Missing values are treated and categorical string values are encoded to integer values as presence of these of feature values prevents the extraction of certain characterization measures.
2. Feature extraction: In this step, the meta-features listed in Tables 4 and 5 are extracted both from the preprocessed data and the original dataset. Table 5 lists the data quality metrics that are proposed by this research for meta-learning. A supporting document is made available in the git repository that explains the formulas/algorithms used to compute all the meta-features.

Dataset characteristics are broadly classified into three dimensions as described in Sect. 2.1. The proposed research takes into account the characteristics of the dataset identified as significant by [23]. Table 4 gives an overview of the direct measures that are considered to model FSDCQ. Meta-features related to data quality are classified into two dimensions. The classification dimension represents the important metrics for machine learning classification tasks.

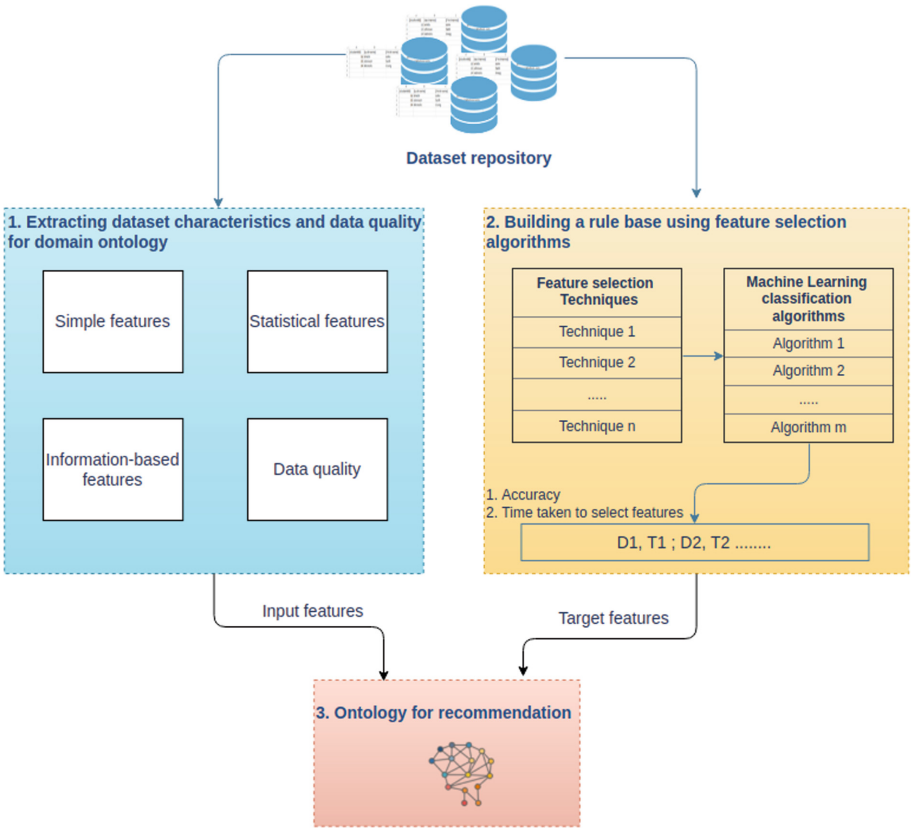


Fig. 2. Proposed recommendation model for feature selection

Intrinsic dimension represents the metrics that are independent of user's context [42]. Table 5 gives a list of data quality metrics that are extracted to model the ontology FSDCQ. The extracted meta-features are populated in the proposed ontology, which is described in Sect. 4.3.

4.2 Building a Rule Base

A rule base is an external knowledge that is added to the ontology to enhance the expressivity of the ontology. This rule base helps to identify the relationship between the feature selection algorithm and the database. Feature selection algorithms are grouped into two broad categories: filter and wrapper. The filter method is based on the dataset characteristics, while the wrapper approach measures the feature subset using the learning algorithm's error rate as the evaluation function. Due to the complex nature of wrapper methods, the proposed research focuses on the filter method for experiments. The proposed study considers a range of feature selection algorithms characterized by their filter classes and evaluation criteria (refer Fig. 3). Feature selection algorithms are evaluated by considering different types of classifiers such as instance (kNN), symbolic (C4.5), statistical (Naive bayes), and connectionist (SVM) approaches. To implement machine learning models, one algorithm is chosen from each type of classifier. Feature selection algorithms for recommendations are ranked based on two performance metrics, 1. Accuracy of the model 2. Time required for the feature selection algorithm to select features. As a result, for each dataset, we have a ranking of the feature selection methods. This ranking is used to determine the optimal feature selection methods, which serve as the target features.

Table 4. Characteristics selected to describe the dataset

Dataset characteristic	Metrics	Description
Simple	Number of classes	Represents the properties taken from the flat file.
	Number of features	
	Number of instances	
Statistical	Average correlation of the feature attributes	Calculates the degree of linear relation degree between all attribute pairs.
	Average asymmetry of the features	Describes the distribution of data from the symmetry condition.
Information	Class entropy	Indicates the probability distribution of observations in a set of data that correspond to a certain class.
	Signal/noise ratio	Indicates the amount of inadequate data in the dataset.
	Equivalent number of attributes	Represents minimum number of attributes required to represent the class

Table 5. Proposed metrics to measure data quality

Dimension	Metrics	Description
Classification	Class overlap	When a region in the data space contains data points from multiple classes.
	Outlier detection	Identifies an unusual data item.
	Class imbalance	Indicates difference in the number of examples in each class. It can be calculated with the entropy of class proportions, imbalance ratio.
Intrinsic	Completeness	Refers to the comprehensiveness or wholeness of the data.
	Conciseness	Refers to uniqueness of the data points.
	Accuracy	Refers to whether the data values stored for an object are the correct values

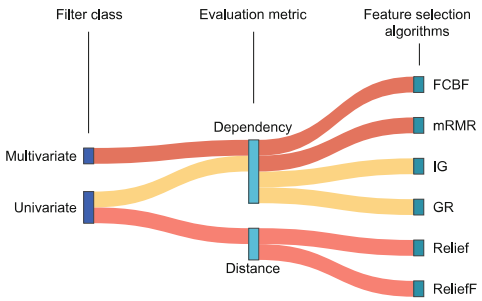


Fig. 3. Feature selection algorithms considered in FSDCQ

4.3 Populating Ontology for the Recommendations

Meta-features that are described in Sect. 4.1 are populated as individuals in the ontology along with highly ranked feature selection algorithms that are calculated in Sect. 4.2. It acts as historical data for recommendations. These meta-features are uplifted using mapping languages. Some of the existing mapping languages are R2RML [16,30], JUMA [10], MappingMaster [18]. Semantic Web Rule Language (SWRL) rules are formulated to recommend feature selection algorithms that are based on historical data. Meta-features will be antecedent of the SWRL rule where as feature selection algorithm will be consequent. Listing 4.1 shows sample of SWRL rule where ?d1 and ?d2 are variables to unify dataset instances, ?mf1 for meta-feature 1, ?fsa for feature selection algorithm. Axiom 'differentFrom' is important to avoid same dataset instances getting binded for variables d1 and d2. SWRL selects feature selection algorithm for dataset d2, if all the attributes of d1 and d2 are same.

```
dcat:dataset(?d1) ^ dcat:dataset(?d2) ^ FSDCQ:hasMF1(?d1, ?mf1)
^ FSDCQ:hasMF1(?d2, ?mf1) ^ FSDCQ:hasFSA(?d1, ?fsa)
^ differentFrom(?d1,?d2) -> sqwrl:select(?d2, ?fsa)
```

Listing 4.1. SWRL rule format for recommendations

5 Experimental Results and Discussion

The overall goal of the FSDCQ is to provide assistance with decision-making phases that affect the result of the knowledge discovery process. It concentrates on two stages of the CRISP-DM process (data understanding and data preparation), which need a significant search for alternative approaches. One such approach is feature selection. Data mining practitioners can consult the FSDCQ ontology to describe meta-features of the dataset. Another application of FSDCQ is meta-learning, which involves the analysis of meta-features to recommend the feature selection algorithm. Thus, the novel objective is to support meta-analysis of machine learning experiments to automatically identify feature selection algorithms that are predictive of good or bad performance. Experiments are conducted on a laptop running Linux Mint 19.3 Cinnamon and powered by an Intel(R) Core(TM) i7-9750H CPU running at 2.60GHz with 16GB of RAM. The experiment is publicly accessible through a git repository[2] and makes use of ten datasets from the UCI repository. Dataset characteristics and quality information are extracted as mentioned in Sect. 4.1. Basic dataset characteristics of the considered dataset are tabulated in Table 6. Datasets are considered to have a small to a large number of features, a small to a large number of attributes, and be a binary or multiclass. Datasets are preprocessed to extract their characteristics and quality information.

Table 6. Basic dataset characteristics

Dataset	Features	Attributes	Classes
Wholesale customer	8	440	2
Caesarian	6	79	2
Bank	17	45211	2
Bank note	5	1371	2
Heart failure	13	299	2
Wine	14	177	3
HCV energy	29	1385	4
Las vegas trip	20	504	7
Iris	5	149	3
Glass	11	213	6

[2] https://github.com/aparnanayakn/onto-DCQ-FS.git.

The classification accuracy of the model and the time required to select features by each feature selection algorithm are used to rank feature selection algorithms for each dataset. However, classification algorithms exhibit varying degrees of bias. In order to overcome this limitation, four representative classification algorithms are considered in the proposed research. Highlighted algorithms in each type are considered for evaluating feature selection techniques.

The extracted characteristics and quality features are mapped to the proposed ontology FSDCQ using MappingMaster [18]. MappingMaster is a domain-specific language for defining spreadsheet-to-OWL ontology mappings. It allows to map individuals to the ontology by mapping classes, object properties, and data properties. The Fig. 4 depicts the screenshot the Protégé after it has been populated with individuals. We can observe that file 'test1.csv' has no feature selection algorithms in property assertions.

Relationships between individuals have to be inferred to recommend a feature selection algorithm. SWRL is a rule-based language that extends the ontology axioms with rules in antecedent-consequent form. These rules are based on OWL classes and properties, which work on the concept of unification. Object properties that describe meta-features will be antecedent of the rules. Corresponding feature selection algorithms will be the consequent of the rules.

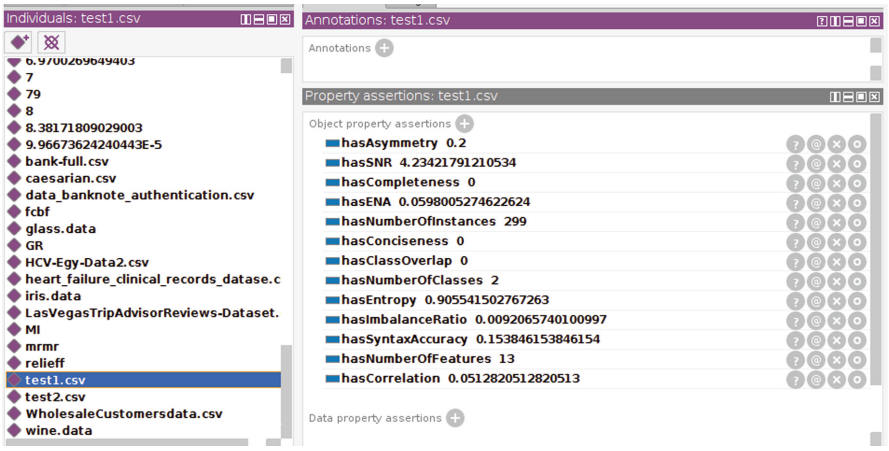


Fig. 4. Individuals and their properties

The proposed work has two key components. First, domain ontology, FSDCQ which can be evaluated using competency questions. Competency questions are answered with the help of SPARQL queries. This helps users understand the domain represented in the ontology. Another key component is the rule-based recommendation model, which can be evaluated using the recommendation hit ratio. This metric is evaluated by comparing the time taken to select features by the recommended feature selection algorithm and accuracy of the classifiers by

incorporating the recommended feature selection algorithm with the accuracy of classifiers with non-recommended feature selection algorithms. However, in the current experiment, ten samples are considered, along with an additional two samples for testing. Recommendations for two testing samples is as shown in Fig. 5. These testing samples have same features to that of testing samples, which can be seen in Fig. 6.

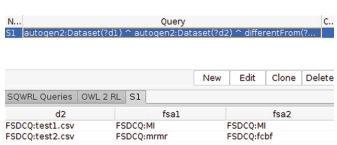
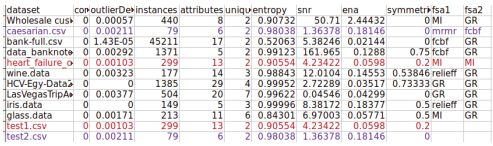


Fig. 5. Recommendations using SQWRL



dataset	cov	outlierDe	instances	attributes	uniqu	entropy	snr	ena	symmetri	fsa1	fsa2
Wholesale cus	0	0.00057	440	8	2	0.90732	50.71	2.44432	0	MI	GR
caesarian.csv	0	0.00211	79	6	2	0.98038	1.36378	0.18146	0	mrmr	fcbf
bank-full.csv	0	1.43E-05	45211	17	2	0.52063	5.38246	0.02144	0	fcbf	GR
data_banknote	0	0.00292	1371	5	2	0.99123	161.965	0.1288	0.75	fcbf	GR
heart_failure_c	0	0.00103	299	13	2	0.90554	4.23422	0.0598	0.2	MI	MI
wine.data	0	0.00323	177	14	3	0.98843	12.0104	0.14553	0.53846	relieff	GR
HCV-Egy-Data	0	0	1385	29	4	0.99952	2.72289	0.03517	0.73333	GR	GR
LasVegasTripA	0	0.00377	504	20	7	0.99622	0.04546	0.04299	0	GR	GR
iris.data	0	0	149	5	3	0.99996	8.38172	0.18377	0.5	relieff	GR
glass.data	0	0.00171	213	11	6	0.84301	6.97003	0.05771	0.5	MI	GR
test1.csv	0	0.00103	299	13	2	0.90554	4.23422	0.0598	0.2		
test2.csv	0	0.00211	79	6	2	0.98038	1.36378	0.18146	0		

Fig. 6. FSDCQ individuals in flat file format

6 Conclusion and Future Works

We introduced the FSDCQ ontology in this research study. It establishes a conceptual framework for meta-learning and the links between meta-features in order to facilitate algorithm recommendation. The methodology proposed for recommending feature selection algorithms establishes relationships between ontology individuals and unifies them to recommend feature selection algorithms.

In a future study, we will strengthen the FSDCQ ontology by enhancing the expressivity of SWRL rules. In the proposed research, the unification property is utilized for the recommendation. However, in the real-world, we may have many situations where multiple features of the dataset are similar but not the same values. Unification fails to recommend feature selection algorithms in such cases. Identifying the most frequently occurring pattern as a recommendation rule will be other future work of the study. Another interesting extension would be clustering the datasets based on their domain, and feature selection algorithm recommendation can be based on the domain. Additionally, FSDCQ can be upgraded to identify the root causes of data quality problems.

Acknowledgements. This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

1. Albertoni, R., Isaac, A.: Introducing the data quality vocabulary (DQV). Semantic Web **12**(1), 81–97 (2021)
2. Almeida, R., Maio, P., Oliveira, P., Barroso, J.: An ontology-based methodology for reusing data cleaning knowledge. In: Proceedings of the International Conference on Knowledge Engineering and Ontology Development (KEOD 2015), pp. 202–211. SciTePress (2015)

3. Bozic, B., Brennan, R., Feeney, K., Mendel-Gleason, G.: Describing reasoning results with RVO, the reasoning violations ontology. In: MEPDaW and LDQ Co-located with ESWC, CEUR Workshop Proceedings, vol. 1585, pp. 62–69 (2016)
4. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Comput. Electr. Eng. **40**(1), 16–28 (2014)
5. Chen, J., Li, K., Rong, H., Bilal, K., Yang, N., Li, K.: A disease diagnosis and treatment recommendation system based on big data mining and cloud computing. Inf. Sci. **435**, 124–149 (2018)
6. Chen, R.C., Huang, Y.H., Bau, C.T., Chen, S.M.: A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection. Expert Syst. Appl. **39**(4), 3995–4006 (2012)
7. Dash, M., Liu, H.: Feature selection for classification. Intell. Data Anal. **1**(1–4), 131–156 (1997)
8. Fernández-López, M., Gómez-Pérez, A., Juristo, N.: Methontology: from ontological art towards ontological engineering (1997)
9. Fürber, C., Hepp, M.: Towards a vocabulary for data quality management in semantic web architectures. In: Proceedings of the 2011 EDBT/ICDT Workshop on Linked Web Data Management, pp. 1–8. ACM (2011)
10. Junior, A.C., Debruyne, C., Longo, L., O'Sullivan, D.: On the mental workload assessment of uplift mapping representations in linked data. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2018. CCIS, vol. 1012, pp. 160–179. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-14273-5_10
11. Kalousis, A., Hilario, M.: Feature selection for meta-learning. In: Cheung, D., Williams, G.J., Li, Q. (eds.) PAKDD 2001. LNCS (LNAI), vol. 2035, pp. 222–233. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-45357-1_26
12. Keet, C.M., Lawrynowicz, A., d'Amato, C., Kalousis, A., Nguyen, P., Palma, R., Stevens, R., Hilario, M.: The data mining optimization ontology. J. Web Semant. **32**, 43–53 (2015)
13. Longo, L., Goebel, R., Lecue, F., Kieseberg, P., Holzinger, A.: Explainable artificial intelligence: concepts, applications, research challenges and visions. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2020. LNCS, vol. 12279, pp. 1–16. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-57321-8_1
14. Mantovani, R.G., Rossi, A.L.D., Alcobaça, E., Vanschoren, J., de Carvalho, A.C.P.L.F.: A meta-learning recommender system for hyperparameter tuning: predicting when tuning improves SVM classifiers. Inf. Sci. **501**, 193–221 (2019)
15. Nakamura, M., Otsuka, A., Kimura, H.: Automatic selection of classification algorithms for non-experts using meta-features. China-USA Bus. Rev. **13**(3) (2014)
16. Nayak, A., Bozic, B., Longo, L.: Extending r2rml-f to support dynamic datatype and language tags. Proc. Comput. Sci. **192**, 709–716 (2021). Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021
17. Obeid, C., Lahoud, I., El Khoury, H., Champin, P.A.: Ontology-based recommender system in higher education. In: Companion Proceedings of the The Web Conference 2018, pp. 1031–1034 (2018)
18. O'Connor, M.J., Halaschek-Wiener, C., Musen, M.A.: Mapping master: a flexible approach for mapping spreadsheets to OWL. In: Patel-Schneider, P.F., et al. (eds.) ISWC 2010. LNCS, vol. 6497, pp. 194–208. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-17749-1_13
19. Oreski, D., Oreski, S., Klicek, B.: Effects of dataset characteristics on the performance of feature selection techniques. Appl. Soft Comput. **52**, 109–119 (2017)

20. Panov, P., Dzeroski, S., Soldatova, L.N.: Ontodm: An ontology of data mining. In: Workshops Proceedings of the 8th IEEE International Conference on Data Mining, pp. 752–760. IEEE Computer Society (2008)
21. Panov, P., Soldatova, L., Džeroski, S.: OntoDM-KDD: ontology for representing the knowledge discovery process. In: Fürnkranz, J., Hüllermeier, E., Higuchi, T. (eds.) DS 2013. LNCS (LNAI), vol. 8140, pp. 126–140. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40897-7_9
22. Panov, P., Soldatova, L.N., Dzeroski, S.: Generic ontology of datatypes. Inf. Sci. **329**, 900–920 (2016)
23. Parmezan, A.R.S., Lee, H.D., Spolaôr, N., Wu, F.C.: Automatic recommendation of feature selection algorithms based on dataset characteristics. Expert Syst. Appl. **185**, 115589 (2021)
24. Parmezan, A.R.S., Lee, H.D., Wu, F.C.: Metalearning for choosing feature selection algorithms in data mining: proposal of a new framework. Expert Syst. Appl. **75**, 1–24 (2017)
25. Peng, Y., Flach, P.A., Soares, C., Brazdil, P.: Improved dataset characterisation for meta-learning. In: Lange, S., Satoh, K., Smith, C.H. (eds.) DS 2002. LNCS, vol. 2534, pp. 141–152. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-36182-0_14
26. Pise, N., Kulkarni, P.: Algorithm selection for classification problems. In: SAI Computing Conference (SAI), pp. 203–211. IEEE (2016)
27. Reif, M., Shafait, F., Dengel, A.: Prediction of classifier training time including parameter optimization. In: Bach, J., Edelkamp, S. (eds.) KI 2011. LNCS (LNAI), vol. 7006, pp. 260–271. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24455-1_25
28. Reif, M., Shafait, F., Goldstein, M., Breuel, T., Dengel, A.: Automatic classifier selection for non-experts. Pattern Anal. Appl. **17**(1), 83–96 (2012). https://doi.org/10.1007/s10044-012-0280-z
29. Rivolli, A., Garcia, L.P., Soares, C., Vanschoren, J., de Carvalho, A.C.: Meta-features for meta-learning. Knowl. Based Syst. **240**, 108101 (2022)
30. Rodriguez-Muro, M., Rezk, M.: Efficient sparql-to-sql with r2rml mappings. J. Web Semant. **33**, 141–169 (2015)
31. Rosa, R.L., Schwartz, G.M., Ruggiero, W.V., Rodríguez, D.Z.: A knowledge-based recommendation system that includes sentiment analysis and deep learning. IEEE Trans. Indust. Inf. **15**(4), 2124–2135 (2018)
32. Shilbayeh, S., Vadera, S.: Feature selection in meta learning framework. In: Science and Information Conference, pp. 269–275. IEEE (2014)
33. Song, Q., Wang, G., Wang, C.: Automatic recommendation of classification algorithms based on dataset characteristics. Pattern Recogn. **45**(7), 2672–2689 (2012)
34. Tianxing, M., Myint, M., Guan, W., Zhukova, N., Mustafin, N.: A hierarchical data mining process ontology. In: 28th Conference of Open Innovations Association (FRUCT), pp. 465–471. IEEE (2021)
35. Uschold, M., Gruninger, M.: Ontologies: principles, methods and applications. Knowl. Eng. Rev. **11**(2), 93–136 (1996)
36. Vanschoren, J.: Meta-learning: A Survey. arXiv preprint arXiv:1810.03548 (2018)
37. Vanschoren, J., Soldatova, L.: Exposé: an ontology for data mining experiments. In: International Workshop on Third Generation Data Mining: Towards Service-Oriented Knowledge Discovery (SoKD-2010), pp. 31–46 (2010)
38. Vilalta, R., Giraud-Carrier, C.G., Brazdil, P., Soares, C.: Using meta-learning to support data mining. Int. J. Comput. Sci. Appl. **1**(1), 31–45 (2004)

39. Vilone, G., Longo, L.: Classification of explainable artificial intelligence methods through their output formats. Mach. Learn. Knowl. Extract. **3**(3), 615–661 (2021)
40. Vilone, G., Longo, L.: Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf. Fusion **76**, 89–106 (2021)
41. Wang, G., Song, Q., Sun, H., Zhang, X., Xu, B., Zhou, Y.: A feature subset selection algorithm automatic recommendation method. J. Artif. Intell. Res. **47**, 1–34 (2013)
42. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: a survey. Semantic Web **7**(1), 63–93 (2016)
43. Zhongguo, Y., Hongqi, L., Ali, S., Yile, A.: Choosing classification algorithms and its optimum parameters based on data set characteristics. J. Comput. **28**(5), 26–38 (2017)