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

The Version of Record of this contribution is published in ICWE 2022. Lecture Notes in Computer Science, vol 13362, and is available online at https://doi.org/10.1007/978-3-031-09917-5_20

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 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 feature selection 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 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 [5, 9, 10]. Inappropriate feature

selection algorithms can cause serious problems, such as compromising the quality of the patterns to be learnt from data and, thus, model performance. A common approach is 'trial-and-error', which tends to be often effective [22]. An alternative is to select a feature selection algorithm based on the characteristics of the dataset. Specifically, this can be implemented by using meta-learning concepts [40] and by utilizing dataset characteristics that are called "meta-features". Automating the selection of feature selection algorithms is a challenge in data mining. However, if overcome, it promises to accelerate the productivity of data scientists and machine learning practitioners [27]. There exists a relationship between the performance of a feature selection algorithm and the characteristics of the dataset [35].

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, which acts as a recommender. The benefits of using an ontology to deliver such a recommendation include interoperability, potential reuse, and sharing of knowledge [39]. 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. In order to achieve this, the work seeks to augment meta-features with data quality information.

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 Section 4. Results of such an experiment are presented and discussed in 5 Finally, Section 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. Although various feature selection algorithms have been proposed, some of these outperform others in terms of performance (for example, classification accuracy) for a given dataset [43]. This

The Version of Record of this contribution is published in ICWE 2022. Lecture Notes in Computer Science, vol 13362, and is available online at https://doi.org/10.1007/978-3-031-09917-5_20

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 [32]. 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 [42]. 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 [1, 33, 37], health-care [6, 7], higher education [20]. 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 [11] 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 ontology for data mining with the aim of providing a unified framework for data mining research. It attempts to cover the full width of the data mining cycle, containing high-level classes, such as

The Version of Record of this contribution is published in ICWE 2022. Lecture Notes in Computer Science, vol 13362, and is available online at https://doi.org/10.1007/978-3-031-09917-5_20

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
[15]	Ranking based on McNemar test	1082*	5	8	Accuracy	✓	✓	✗	✗
[18]	SVM	156	-	7	Accuracy	✓	✓	✗	✗
[19]	kNN	58	-	-	F1 score				
[22]	C5.0 decision tree	128	5	3	Accuracy, time complexity	✓	✓	✗	✗
[26]	Ranking based on MCPM	213	5	5	Learning time, Percentage of selected attributes, Error rate	✓	✓	✓	✓
[28]	kNN	47	-	10	Spearman's rank correlation	✓	✓	✗	✓
[29]	kNN	38	-	9	Accuracy	✓	✓	✗	✗
[30]	Regression	123	-	5	Correlation	✓	✓	✓	✗
[31]	Regression	54	-	9	Accuracy	✓	✓	✓	✓
[35]	J4.8 decision trees	26	4	3	Accuracy	✓	✗	✗	✗
[36]	kNN	84	-	-	Accuracy, Execution time	✓	✗	✗	✗
[43]	kNN	115	22	5	Recommendation hit ration based on accuracy	✓	✓	✗	✗
[45]	Variance, LIBSVM	84	-	3	Accuracy	✓	✓	✓	✓

The Version of Record of this contribution is published in ICWE 2022. Lecture Notes in Computer Science, vol 13362, and is available online at https://doi.org/10.1007/978-3-031-09917-5_20

data mining tasks and algorithms, and more specific classes related to certain sub-fields, such as constraints [23]. 'Expose' is an ontology to describe machine learning experiments in a standardised fashion. This ontology is used to express and share experiment meta data [41].

To represent the relationship between data mining tasks and dataset characteristics, multiple ontologies have been designed. 'OntoDM-KDD' [24], 'OntoDT' [25], 'CRISP-DM' [38] 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 [16]. 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 a vocabulary that describes the conceptualization of a domain, supporting standardized formulation of data quality, cleansing of rules, classification of data quality problems, and the computation of data quality scores [12]. Data Cleaning Ontology (DCO) refines and extends data cleaning operations which directly assesses data quality [3]. Another matured ontology is recommended by the World Wide Web Consortium (W3C)[1] which covers most of the aspects of data quality [2].

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 reusing classes from the 'OntoDT', 'OntoDM-KDD', 'CRISP-DM' ontologies along with the 'DCO', 'DQM', and 'DQV' ontologies. The W3C recommendation ontology language, OWL (Web Ontology Language), is adopted to develop such an ontology with Protégé [2] 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 Table 4 and 5. The class hierarchy of the FSDCQ ontology is shown in Figure 1. Table 2

[1] https://www.w3.org/TR/vocab-dqv/
[2] https://protege.stanford.edu/

The Version of Record of this contribution is published in ICWE 2022. Lecture Notes in Computer Science, vol 13362, and is available online at https://doi.org/10.1007/978-3-031-09917-5_20

depicts ontology metrics of FSDCQ before adding individuals.

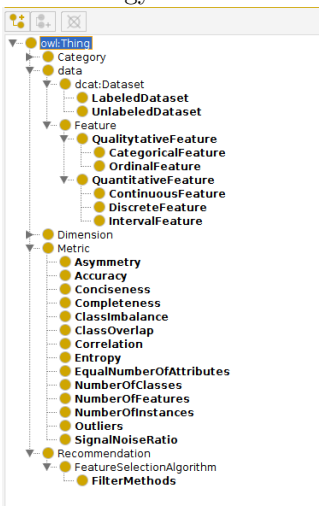

Fig. 1: Class hierarchy of FSDCQ

Table 2: FSDCQ metrics

Property	Count
Axioms	396
Classes	39
Logical axioms	326

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.

Competency questions are directed at users and help us define the scope of an ontology. In other words, they are the questions to which users seek answers by exploring and querying an ontology and its associated knowledge base. Specifically, the principal competency questions associated to an FSDCQ are:

- **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.

The Version of Record of this contribution is published in ICWE 2022. Lecture Notes in Computer Science, vol 13362, and is available online at https://doi.org/10.1007/978-3-031-09917-5_20

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 is the set of dataset characteristics required for a feature selection algorithm X?
CQ6: Which algorithm should be used (or avoided) when a dataset has many more variables than instances?

4 Proposed methodology

This section presents a recommendation model for feature selection algorithm, as depicted in Figure 2). It consists of the following three major stages of implementation:

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

These stages are described in details in the following sections.

4.1 Extraction of dataset characteristics and quality

Dataset repository consists of multiple datasets, which will be used to extract meta-features. We represent each dataset with - rows and - features in the flat file format.

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 Table 4 and Table 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 available at the link [3] that explains the formulas/algorithms used to compute all the meta-features.

[3] `https://github.com/aparnanayakn/onto-DCQ-FS/blob/main/Supporting%20document_FSDCQ.pdf`

The Version of Record of this contribution is published in ICWE 2022. Lecture Notes in Computer Science, vol 13362, and is available online at https://doi.org/10.1007/978-3-031-09917-5_20

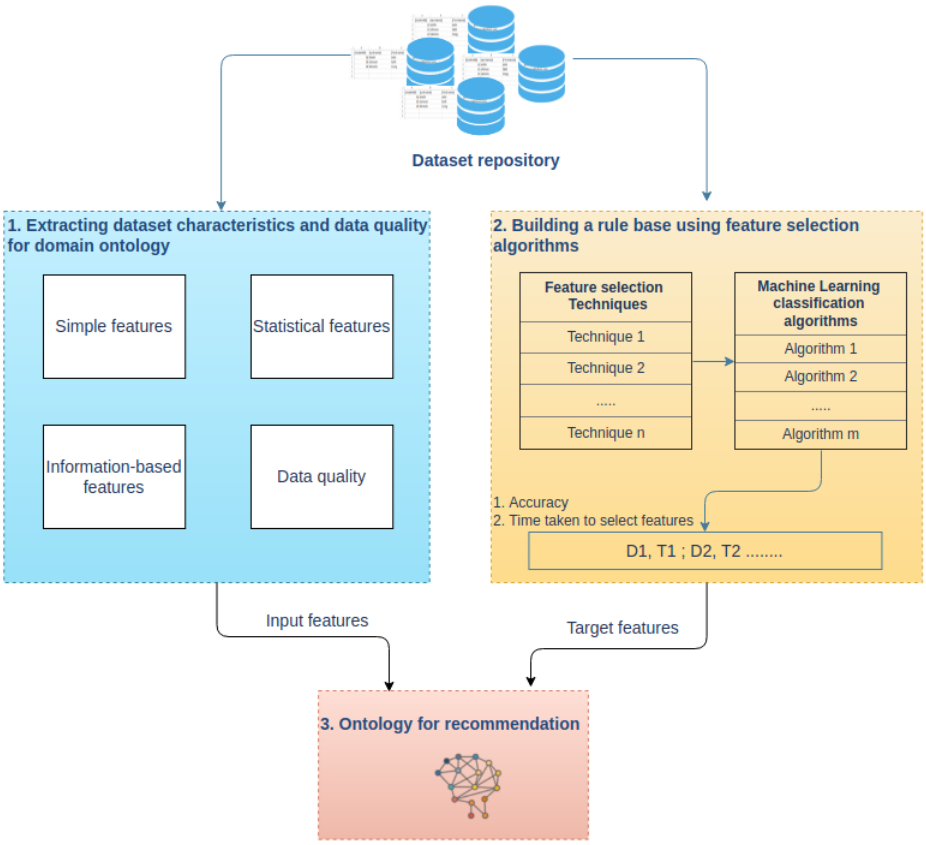

Fig. 2: Proposed recommendation model for feature selection

Dataset characteristics are broadly classified into three dimensions as described in Section 2.1. The proposed research takes into account the characteristics of the dataset identified as significant by [26]. 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. Intrinsic dimension represents the metrics that are independent of user's context [44]. 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 Section 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

The Version of Record of this contribution is published in ICWE 2022. Lecture Notes in Computer Science, vol 13362, and is available online at https://doi.org/10.1007/978-3-031-09917-5_20

Table 4: Characteristics selected to describe the dataset

Dataset characteristic	**Metrics**	**Description**
Simple	Number of classes	Depict properties taken from the attribute-value table
	Number of features	
	Number of instances	
Statistical	Average correlation of the feature attributes	Measures the linear relation degree between random attribute pairs.
	Average asymmetry of the features	Describes how and how much the data distribution departs from the symmetry condition.
Information	Class entropy	Reflects the approximate amount of information required to identify the class of an example from the dataset.
	Signal/noise ratio	Expresses the amount of non-useful information of a dataset
	Equivalent number of attributes	Ratio between the entropy of the class and the average mutual information between classes and attribute

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	A large difference in the number of examples per class in the training dataset. It can be computed using 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.

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 method uses the error rate of the learning algorithm as the evaluation function

The Version of Record of this contribution is published in ICWE 2022. Lecture Notes in Computer Science, vol 13362, and is available online at https://doi.org/10.1007/978-3-031-09917-5_20

to measure the feature subset. Due to the complex nature of wrapper methods, the proposed research focuses on the filter method for experiments. The proposed work takes into account a variety of feature selection algorithms that are classified according to their filter classes and evaluation criteria (refer Figure 3). Feature selection algorithms are evaluated by considering different types of classifiers (refer Table 6). 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, we have a ranking of the feature selection algorithms for each dataset. We incorporate this ranking to identify the best feature selection algorithms, which act as the target features.

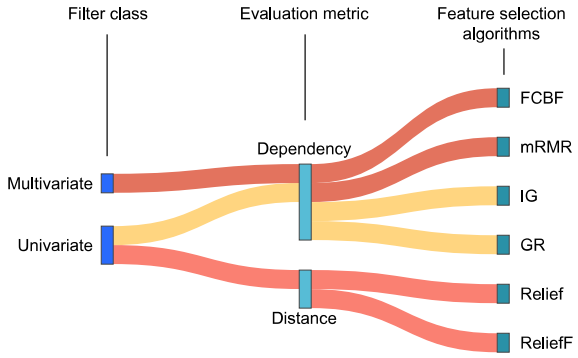

Fig. 3: Feature selection algorithms considered in FSDCQ

Table 6: Classification algorithms considered to evaluate feature selection algorithms

Type	**Algorithms**
Instance	**k-Nearest Neighbour (kNN)** [14], Incremental Hypersphere Classifier (IHC) [17]
Symbolic(Rule)	**C4.5** [34], PRISM [4], RIPPER [8]
Statistical	**Naive bayes** [13]
Connectionist	**SVM**, ANN

The Version of Record of this contribution is published in ICWE 2022. Lecture Notes in Computer Science, vol 13362, and is available online at https://doi.org/10.1007/978-3-031-09917-5_20

4.3 Populating ontology for the recommendations

Meta-features that are described in Section 4.1 are populated as individuals in the ontology along with highly ranked feature selection algorithms that are calculated in Section 4.2. It acts as historical data for recommendations. 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 support for decision-making steps that impact the outcome of the knowledge discovery process. It focuses on two phases of the CRISP-DM process (data understanding and data preparation), which demand a non-trivial search in the space of alternative methods. One such process is the selection of features. 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[4] and makes use of ten datasets from the UCI repository. Dataset characteristics and quality information are extracted as mentioned in Section 4.1. Basic dataset characteristics of the considered dataset are tabulated in Table 7. 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.

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

The Version of Record of this contribution is published in ICWE 2022. Lecture Notes in Computer Science, vol 13362, and is available online at https://doi.org/10.1007/978-3-031-09917-5_20

Table 7: Basic dataset characteristics

Dataset	**Number of features**	**Number of attributes**	**Number of 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

To rank feature selection algorithms for each dataset, the classification accuracy of the model and the time required to select features by each feature selection algorithm are used. However, classification algorithms exhibit varying degrees of bias. In order to overcome this limitation, four representative classification algorithms are considered in the proposed research. Table 6 lists various algorithms based on instance, symbolic, statistical, and connectionist approaches. 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 [21]. 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 Figure 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.

The Version of Record of this contribution is published in ICWE 2022. Lecture Notes in Computer Science, vol 13362, and is available online at https://doi.org/10.1007/978-3-031-09917-5_20

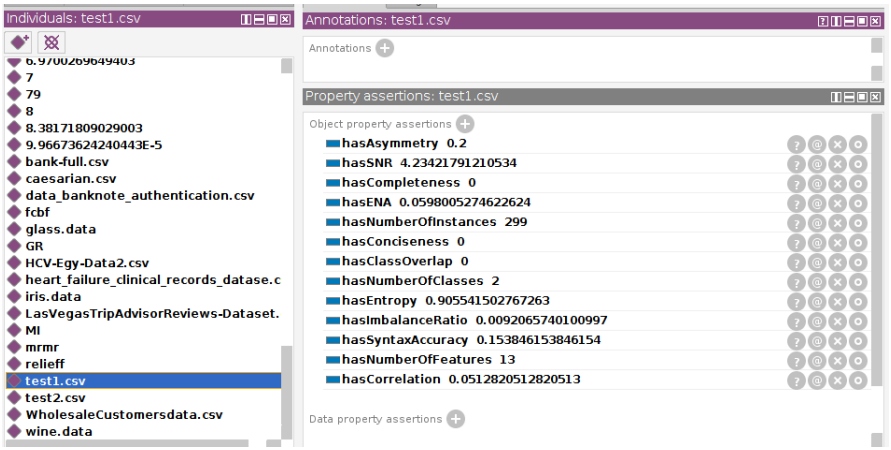

Fig. 4: Individuals and their properties

5.1 Evaluation

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 Figure 5. These testing samples have same features to that of testing samples, which can be seen in Figure 6.

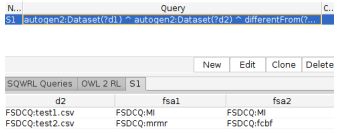

Fig. 5: Recommendations using SQWRL

dataset	co	outlierDe	instances	attributes	uniq	entropy	snr	ena	symmetr	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_banknot	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

In this research work, we have presented the FSDCQ ontology. It provides a conceptual framework for meta-learning and the relationships between

The Version of Record of this contribution is published in ICWE 2022. Lecture Notes in Computer Science, vol 13362, and is available online at https://doi.org/10.1007/978-3-031-09917-5_20

meta-features to enable the recommendation of feature selection algorithms. 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. Aciar, S., Zhang, D., Simoff, S., Debenham, J.: Informed recommender: Basing recommendations on consumer product reviews. IEEE Intelligent systems **22**(3), 39–47 (2007)
2. Albertoni, R., Isaac, A.: Introducing the data quality vocabulary (DQV). Semantic Web **12**(1), 81–97 (2021)
3. Almeida, R., Maio, P., Oliveira, P., Barroso, J.: An ontology-based methodology for reusing data cleaning knowledge. In: KEOD 2015 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development. pp. 202–211. SciTePress (2015)
4. Bramer, M.: Automatic induction of classification rules from examples using n-prism. In: Research and development in intelligent systems XVI, pp. 99–121. Springer (2000)
5. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Computers & Electrical Engineering **40**(1), 16–28 (2014)
6. 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. Information Sciences **435**, 124–149 (2018)
7. 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 Systems with Applications **39**(4), 3995–4006 (2012)
8. Cohen, W.W.: Fast effective rule induction. In: Machine learning proceedings 1995, pp. 115–123. Elsevier (1995)

The Version of Record of this contribution is published in ICWE 2022. Lecture Notes in Computer Science, vol 13362, and is available online at https://doi.org/10.1007/978-3-031-09917-5_20

9. Cunningham, P., Kathirgamanathan, B., Delany, S.J.: Feature selection tutorial with python examples. arXiv preprint arXiv:2106.06437 (2021)
10. Dash, M., Liu, H.: Feature selection for classification. Intelligent data analysis **1**(1-4), 131–156 (1997)
11. Fernández-López, M., Gómez-Pérez, A., Juristo, N.: Methontology: from ontological art towards ontological engineering (1997)
12. 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)
13. van der Gaag, L.C., Capotorti, A.: Naive bayesian classifiers with extreme probability features. In: International Conference on Probabilistic Graphical Models. Proceedings of Machine Learning Research, vol. 72, pp. 499–510. PMLR (2018)
14. Hendrickx, I., van den Bosch, A.: Hybrid algorithms with instance-based classification. In: Machine Learning: ECML 2005, 16th European Conference on Machine Learning. Lecture Notes in Computer Science, vol. 3720, pp. 158–169. Springer (2005)
15. Kalousis, A., Hilario, M.: Feature selection for meta-learning. In: Knowledge Discovery and Data Mining - PAKDD. Lecture Notes in Computer Science, vol. 2035, pp. 222–233. Springer (2001)
16. Keet, C.M., Lawrynowicz, A., d'Amato, C., Kalousis, A., Nguyen, P., Palma, R., Stevens, R., Hilario, M.: The data mining optimization ontology. Journal of web semantics **32**, 43–53 (2015)
17. Lopes, N., Ribeiro, B.: On the impact of distance metrics in instance-based learning algorithms. In: Pattern Recognition and Image Analysis. Lecture Notes in Computer Science, vol. 9117, pp. 48–56. Springer (2015)
18. 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. Information Sciences **501**, 193–221 (2019)
19. Nakamura, M., Otsuka, A., Kimura, H.: Automatic selection of classification algorithms for non-experts using meta-features. China-USA Business Review **13**(3) (2014)
20. 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)
21. O'Connor, M.J., Halaschek-Wiener, C., Musen, M.A.: Mapping master: A flexible approach for mapping spreadsheets to OWL. In: The Semantic Web - ISWC 2010. Lecture Notes in Computer Science, vol. 6497, pp. 194–208. Springer (2010)
22. Oreski, D., Oreski, S., Klicek, B.: Effects of dataset characteristics on the performance of feature selection techniques. Applied Soft Computing **52**, 109–119 (2017)
23. 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)
24. Panov, P., Soldatova, L.N., Dzeroski, S.: Ontodm-kdd: Ontology for representing the knowledge discovery process. In: Discovery Science - 16th International Conference, DS. Lecture Notes in Computer Science, vol. 8140, pp. 126–140. Springer (2013)
25. Panov, P., Soldatova, L.N., Dzeroski, S.: Generic ontology of datatypes. Information Sciences **329**, 900–920 (2016)

The Version of Record of this contribution is published in ICWE 2022. Lecture Notes in Computer Science, vol 13362, and is available online at https://doi.org/10.1007/978-3-031-09917-5_20

26. Parmezan, A.R.S., Lee, H.D., Spolaôr, N., Wu, F.C.: Automatic recommendation of feature selection algorithms based on dataset characteristics. Expert Systems with Applications **185**, 115589 (2021)
27. 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 Systems with Applications **75**, 1–24 (2017)
28. Peng, Y., Flach, P.A., Soares, C., Brazdil, P.: Improved dataset characterisation for meta-learning. In: Discovery Science, 5th International Conference. Lecture Notes in Computer Science, vol. 2534, pp. 141–152. Springer (2002)
29. Pise, N., Kulkarni, P.: Algorithm selection for classification problems. In: SAI Computing Conference (SAI). pp. 203–211. IEEE (2016)
30. Reif, M., Shafait, F., Dengel, A.: Prediction of classifier training time including parameter optimization. In: Advances in Artificial Intelligence. Lecture Notes in Computer Science, vol. 7006, pp. 260–271. Springer (2011)
31. Reif, M., Shafait, F., Goldstein, M., Breuel, T.M., Dengel, A.: Automatic classifier selection for non-experts. Pattern Analysis and Applications **17**(1), 83–96 (2014)
32. Rivolli, A., Garcia, L.P., Soares, C., Vanschoren, J., de Carvalho, A.C.: Meta-features for meta-learning. Knowledge-Based Systems **240**, 108101 (2022)
33. 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 Transactions on Industrial Informatics **15**(4), 2124–2135 (2018)
34. Ruggieri, S.: Efficient c4.5 [classification algorithm]. IEEE transactions on knowledge and data engineering **14**(2), 438–444 (2002)
35. Shilbayeh, S., Vadera, S.: Feature selection in meta learning framework. In: Science and Information Conference. pp. 269–275. IEEE (2014)
36. Song, Q., Wang, G., Wang, C.: Automatic recommendation of classification algorithms based on dataset characteristics. Pattern Recognition **45**(7), 2672–2689 (2012)
37. Sulthana, A.R., Ramasamy, S.: Ontology and context based recommendation system using neuro-fuzzy classification. Computers & Electrical Engineering **74**, 498–510 (2019)
38. 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)
39. Uschold, M., Gruninger, M.: Ontologies: Principles, methods and applications. The knowledge engineering review **11**(2), 93–136 (1996)
40. Vanschoren, J.: Meta-learning: A survey. arXiv preprint arXiv:1810.03548 (2018)
41. 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)
42. Vilalta, R., Giraud-Carrier, C.G., Brazdil, P., Soares, C.: Using meta-learning to support data mining. International Journal of Computer Science Applications **1**(1), 31–45 (2004)
43. Wang, G., Song, Q., Sun, H., Zhang, X., Xu, B., Zhou, Y.: A feature subset selection algorithm automatic recommendation method. Journal of Artificial Intelligence Research **47**, 1–34 (2013)
44. 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)
45. Zhongguo, Y., Hongqi, L., Ali, S., Yile, A.: Choosing classification algorithms and its optimum parameters based on data set characteristics. Journal of Computers **28**(5), 26–38 (2017)

The Version of Record of this contribution is published in ICWE 2022. Lecture Notes in Computer Science, vol 13362, and is available online at https://doi.org/10.1007/978-3-031-09917-5_20