An Ontology of Machine Learning Algorithms for Human Activity Data Processing

Man Tianxing
ITMO University
St. Petersburg, Russia
mantx626@gmail.com

Nataly Zhukova
ITMO University
St. Petersburg, Russia
nazhukova@mail.ru

***Abstract*—Machine learning algorithms are the main tools in the field of data analysis. However, extracting knowledge from data sets originating in real life requires complex data processing. Obtaining the available tidy data sets and selecting the appropriate analysis algorithm are important issues for data analysts. Because of the complexity of the dataset and the diversity of the algorithms the researchers take too much time in selecting and comparing these algorithms. Human Activity Recognition is a typical example in Internet of Things. Its principle is to identify human behavior by analyzing the coordinate data from the sensors on the human body so that we can achieve remote monitoring. A precise Human Activity Recognition application can serve as a real-time monitoring of the elderly or vulnerable behavior. However, due to the unpredictability of human behavior, these sensor data require relatively complex processing. Therefore, we propose an ontology-based algorithm recommendation system. It consists of several parts: algorithm pool, data features, model features, and mathematical theory. The framework provides data researchers with reasonable solutions based on the characteristics of the data set and the task requirements. Especially for the Internet of Things data such as Human Activity Recognition data set, its recommendations can save users much time for analysis and comparison.**

I. Introduction

Due to the popularization of Internet, A lot of relevant data are generating from human daily activities. Machine learning algorithms are the most effective tools for the conversion from data to knowledge. Experts enhance and improve machine learning and data processing technologies. This causes the confusion about how to choose the right algorithm or technology for data processing for researchers.

However, it is obvious that taxonomies can't include all the complex algorithms and techniques. Therefore, we also strive to propose a framework to be an extensible and portable system so that it is gradually being improved in using.

In response to such requests, Ontology technology becomes our best choice for system construction. Its main advantage is to make the information on the Web that can be understood easily by a computer, realize the semantic interoperation among the information systems with the support of the ontology and intelligently access and retrieve Web resources.

Internet of Things technology is an important part of the new generation of information technology. Due to coming from real life, these data have more uncertainty. They need to go through a complex preprocessing, and then be applied with the appropriate data analysis algorithms to extract useful information. Therefore, our data processing recommendation framework can be applied in the field of IoT.

Human Activities Recognition is a typical application example in Internet of Things. We try to apply our system on it. Experiments show that this system provides effective recommendations for such data processing. And based on the experimental results, the processing provided by our framework gains a high classification accuracy.

The rest of this paper is organized as follow: Section 2 describes relevant knowledge involved in this paper. Section 3 presents the composition and workflow of this ontology-based algorithm recommendation system. Section 4 presents the application of the proposed system on the Human Activity Recognition dataset and compares it with the results of other unrecommended algorithms. Section 5 presents the main conclusion and points directions for future work.

II. Knowledge

A. Ontology

The Semantic Web is an emerging concept that is an intelligent network that can make judgments based on semantics to enable unobstructed communication between people and computers. Each computer connected to the Semantic Web not only understands words and concepts, but also understands the logical relationship between them. Ontology is such a conceptual modeling tool that describes information systems at the semantic and knowledge level. The goal is to capture knowledge in related fields, identify commonly recognized terms in this field, describe the semantics of concepts through the relationships between concepts, and provide a common understanding of the field knowledge [9].

B. Taxonomy of Machine Learning algorithms

At present, A lot of taxonomies of ML algorithms have been proposed. Some researches that are specifically tailored to choose the best performing technology are spreading. These taxonomies are mainly based on the type of algorithm. However, in practice, such taxonomies do not provide users with a valid choice suggestion. We are going to enrich the data processing library on the basis of these existing taxonomies and

define more relationships to create an ontology of machine learning algorithms [10][12].

C. Human Activities Recognition

Human Activities Recognition is a typical example of IoT applications. It uses the sensors on smartphones that people often bring around to receive coordinate data and obtain human actions by analyzing the changes in coordinates and accelerations.

At present, many research results and methods have been proposed. The main application areas are the real-time monitoring of the elderly or vulnerable and the analysis and recording of the physical status [3][7][8]. The popular technologies include deep convolutional neural networks [5], support vector machine [6], Hidden Markov Model [4], and Dynamic Bayesian Networks. Most of them can give good performance in activity recognition.

III. A FRAMEWORK OF MACHINE LEARNING ALGORITHMS

A. Relationships

In usual taxonomy only a "has-a" relationship is existing which can only express the algorithm belongs to a category. This kind of taxonomy does not give the user a suggestion about algorithm selection. In this regard, we also define some necessary Object properties in our ontology:

i. "isFrom": It connects the algorithm and the algorithm source.

ii. "BeSuitableFor": This is one of the main relationships in this framework. Because it connects the dataset feature and the algorithms that have good performance for the feature and the preprocessing techniques that can solve the problem (you can also connect dataset features and some suitable mathematical functions such as Distance Function).

iii. "Apply" "BeAppliedOn": This is a pair of Inverse functional relationship. As a straight-back connection it links algorithm and the related mathematical knowledge.

iv. "Generate" "GeneratedFrom": This is also a pair of Inverse functional relationship which Connects algorithm and the features of algorithm generated model.

B. Basic Structures

The ultimate goal of our framework is to provide effective data processing advice for any data set processing tasks. So, except algorithm pool all the other ontologies that are integrated in this framework serve this purpose. Here are the main components in our framework:

1) Algorithm pool: It is a set of machine learning algorithms and preprocessing technologies. It is based on the DBpedia "Classification Algorithms" entry. We have added some of the missing algorithms in the entry. The other part of the algorithm pool is data preprocessing technology.

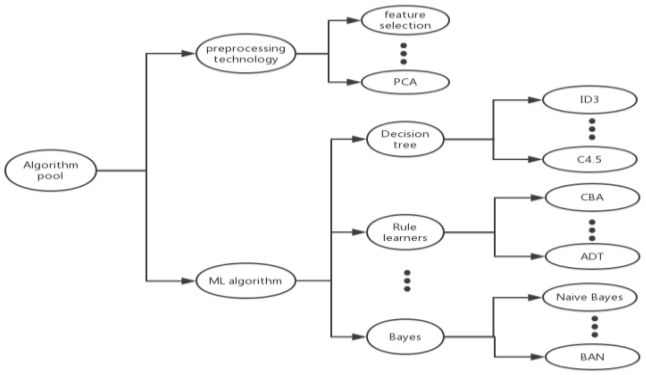

Fig. 1. Main part of the algorithm pool

2) Dataset features: The basic characteristics of a dataset are the decisive factors for selecting the most suitable algorithm in data processing process. Mainly include the following aspects:

i. characters of data samples

ii. characteristics of data features

Fig. 2. Main part of data features

3) Mathematics: Another criterion for choosing a suitable algorithm is the mathematical model in the algorithm. There are mainly such theories:

i. Linear algebra

ii. Probability and Statistics

iii. Multivariable Calculus

iv. Algorithms and Complex Optimization

v. Other: Some other math topics are not covered in the four main areas above. These topics include real and complex analysis, information theory, function spaces and number sets.

We create such an ontology of these mathematical knowledge. This is a good description system for some users who do not understand the theory of machine learning.

4) Intermediate model features: The requirements for the output classification model are another important factor for selecting the ML algorithm.

C. Workflow

The ontologies of data features, mathematical theory, output model features and algorithm pools should be linked together with some specific relationship.

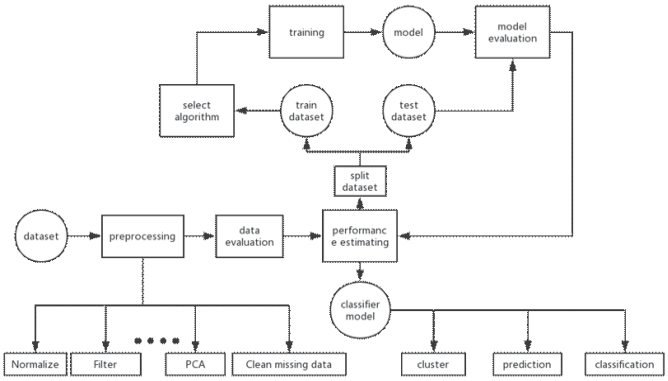

Fig. 3. The process of ML algorithm

When we receive a task to process dataset, we first need to evaluate it to get the data defects which should be solved. Based on these features we find a suitable technology in the algorithm pool for data preprocessing. Every time fixing a dataset defect the dataset is transformed to be a new form. Then another data evaluation is in progress. In each operation we obtain a new dataset form the dataset become better and better until we get a clear enough dataset. Then the task goes to the next step-data analysis. We can select ML algorithm according to dataset feature, the mathematical theory model and the characteristics of output model in the framework [2][11][13].

Algorithm 1 Workflow

Parameters:

input dataset *data*

Algorithm pool including ML algorithms and preprocessing technology *A{a1,a2…an;t1,t2…tn}*

evaluation module *e()*

Search technology or algorithm module *S_tech(),S_algo()*

```
1: e(data)
//preprocessing
2: while (data is not tidy enough) do
3:      F_data{f1,f2…fi}←e(data)
4:      tech←S_tech(F_data, A)
5:      data←tech(data)
//data analysis
6: F_task{t1,t2…tj}←e(task)
7: algo←S_algo(F_data, F_task,A)
8: return algo(data)
```

In this way the final output result is actually a reasonable data analysis process. The original dataset goes through a continuous conversion process-from disorderly to gradually regular. At last a ML algorithm is applied to train the clear dataset to generate an efficient classifier. And it can provide users with relevant data analysis conclusions.

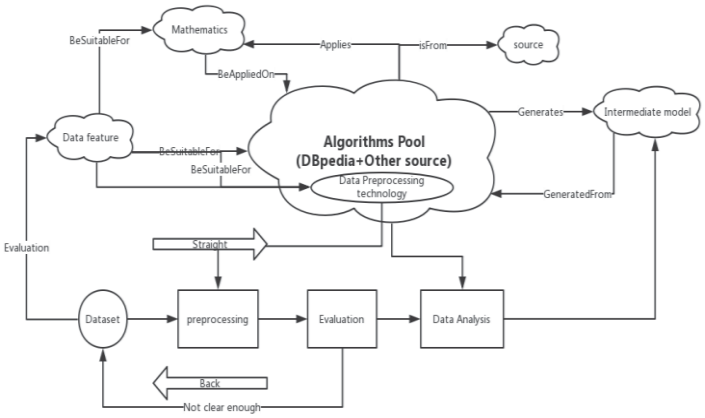

Fig. 4. Workflow

IV. Human Activities Recognition dataset intelligent processing

A. Existing research

As a typical application example of Internet of Things, Human Activities Recognition has attracted the attention of many researchers. But people still rely on their own subjective judgment in dealing with this type of data. There is no systematic scientific evidence to justify these choices. We investigated related research experiments. In the choice of HAR data analysis, people applied a variety of classification algorithms in this project. And the result is good and bad.

In this regard, we selected the more popular experimental data "UCI HAR data set" for experimental verification. Based on the basic characteristics and mission requirements of this data, our recommendation system gradually generates data processing recommendations. We follow this advice to process the data and compare the results with those of other non-selected algorithms.

B. Data preprocessing

Our experimental data is "UCI HAR dataset" from a common database "UCI machine learning repository" [1]. Since it is commonly used experimental data, it is neater than the data set from real life. However, complex preprocessing is still required.

First, the experimental data we obtained came from multiple document files. Data merge is the first step in data preprocessing.

Then the recommended system evaluates this data again. Data set has too many labels. Therefore, we have to do feature selection. Only the labels that are useful for Human Activities Recognition are left. We keep the coordinate movement data and the instantaneous acceleration data.

Then check the missing value and noisy value in the data set and execute delete operation. At this time the data set is very tidy.

Due to the current data is difficult to be analyzed directly. We perform feature calculations to obtain means, medians, MinMax, standard deviations, correlations, etc.

Since we change the labels in this step, the data set has too many labels again. However, we can't extract features by label meanings. According to the system proposal, the best choice is PCA.

After completing the information extraction work, the dataset has become tidy enough for analyzing. The entire preprocessing is shown in Fig 7.

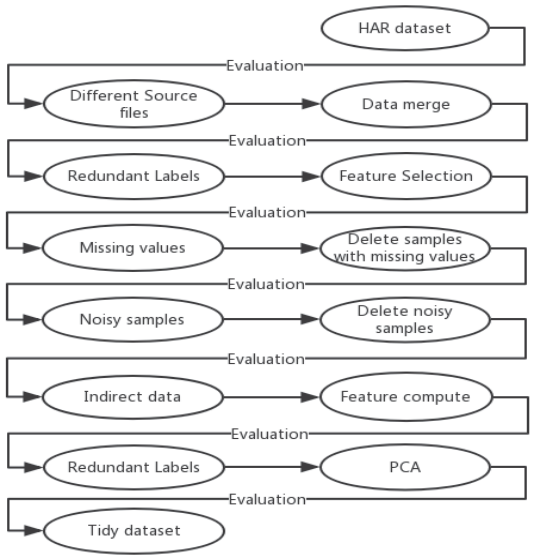

Fig. 5. HAR dataset preprocessing

C. Data analysis

At the stage of data set analysis, choosing a suitable algorithm is the basis of the task. HAR datasets come from a variety of sources so that the characteristics of these datasets are not the same. We cannot get good performance on all datasets with just one analysis algorithm. Therefore, we need to make algorithm selection based on the characteristics of the acquired data set. At the same time the task requirements are another important factor in algorithm selection.

After acquiring a HAR dataset that is sufficiently tidy we will again evaluate the dataset. Then based on the traits obtained from the evaluation, we can find the suitable algorithm in our ontology

Since the current data features are all calculated from the original features, they are Highly interdependent now. The recommended algorithms are: Artificial neural network, Support vector machine.

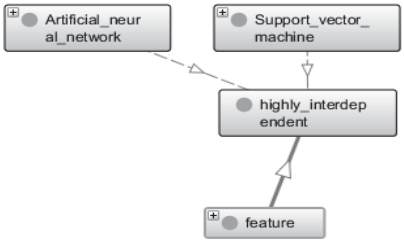

Fig. 6. Query based on 'highly interdependent'

Although we performed feature extraction, respect to 15000 samples dozens of features are still redundant. The recommended algorithms are: Support vector machine.

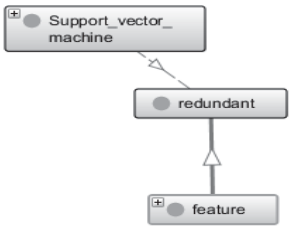

Fig. 7. Query based on 'redundant'

This dataset has only nearly15000 samples, so it belongs to a small-size data set. The applicable algorithm is: Support vector machine.

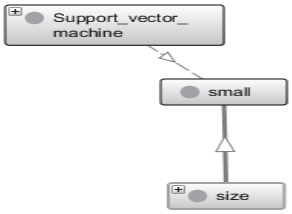

Fig. 8. Query based on 'small size'

In addition, we can make selection depending on task requirements. For Human Activity Recognition the accuracy of classification is the first requirement among all the performances. According to our ontology query the recommended algorithms are: Artificial neural network, Support vector machine, K-nearest neighbors.

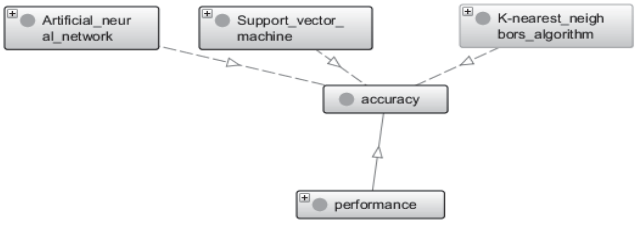

Fig. 9. Query based on 'accuracy'

In summary, the Support vector machine algorithm is the best choice for this task based on the recommendations of our ontology. Because we are doing classification tasks, finally we decide to use Support vector classification algorithm (SVC).

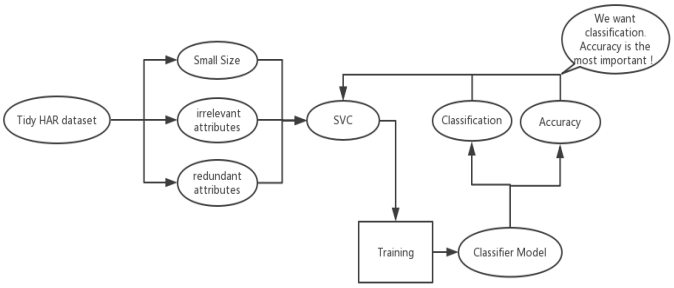

Fig. 10. The process of algorithm selection

In this query, the relationship we searched was 'BeSuitableFor'. If we don't want to miss more choices, we can query the relationship "BeAvailableFor" to get more available algorithms.

D. Result

According to the suggestion of the framework, we cleaned the original dataset. And split it to training dataset and test dataset. SVC algorithm is applied on training dataset for the classification model. At last we use the test dataset verify performance of classification model. The result is as follow (NO. 1-6 corresponding different human activities: 'Walking', 'Walking upstairs', 'Walking downstairs', 'Sitting', 'Standing', 'Laying'; The main performance indicator is: precision, recall and f1-score):

TABLE I. THE RESULT OF SVC

	precision	recall	f1-score	support
1	0.74	0.68	0.71	24419
2	0.81	0.92	0.86	23342
3	0.76	0.83	0.79	21593
4	0.88	0.80	0.84	25336
5	1.00	1.00	1.00	27621
6	0.93	0.90	0.91	27373
avg/total	0.86	0.86	0.86	149684

E. Compare and Conclusion

In order to validate the performance of the framework, we selected some other algorithms used in other Human Activities Recognition studies to do the same classification problem [14][15]. The most common are Naïve Bayes and Decision trees algorithms. The result is as follows:

TABLE II THE RESULT OF NAÏVE BAYES

	precision	recall	f1-score	support
1	1.00	0.53	0.69	24419
2	0.84	1.00	0.91	23342
3	0.75	1.00	0.86	21593
4	0.67	0.96	0.79	25336
5	1.00	0.74	0.85	27621
6	0.86	0.74	0.80	27373
avg/total	0.86	0.82	0.81	149684

TABLE III THE RESULT OF DECISION TREE

	precision	recall	f1-score	support
1	0.74	0.46	0.57	24419
2	0.66	0.92	0.76	23342
3	0.55	0.67	0.60	21593
4	0.73	0.72	0.73	25336
5	0.83	0.83	0.83	27621
6	0.75	0.64	0.69	27373
avg/total	0.72	0.71	0.70	149684

By comparison we can find the overall classification accuracy: SVC> Naïve Bayes> Decision Trees. Of course, many researchers chose SVC algorithm when doing research on Human Activities Recognition. However, this still proves that our framework offers a suitable option.

V. CONCLUSION

Obviously, an ontology of machine learning algorithm has better flexibility than a taxonomy. With the support of ontology technology, we can flexibly define more relationships in the framework. Such a framework effectively helps those non-computer science researchers choose the appropriate method of data processing. Especially when we apply the framework on IoT whose data is from real life, it can provide useful and reasonable advice for our work.

This machine learning algorithm framework is also constantly improving. We are currently focus on a system that automatically evaluates dataset features and how to classify new algorithms based on experimental results.

ACKNOWLEDGEMENT

I would like to express my gratitude to all teachers and classmates helped me during the writing of this thesis. I acknowledge the help of Department of International Postgraduate and Doctoral Studies of ITMO University. They offer me this precious opportunity. I also give a special debt of gratitude to China scholarship council who support my study in Russia. I should finally like to express my gratitude to my beloved parents who have always been helping me out of difficulties and supporting me without a word of complaint.

REFERENCES

[1] UCI machine learning repository, Human Activity Recognition Using Smartphones Data Set, Web: https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones.

[2]Reif, Matthias, et al. "Automatic classifier selection for non-experts." Pattern Analysis and Applications 17.1 (2014): 83-96.

[3]Murata, Satoshi, Masanori Suzuki, and Kaori Fujinami. "A wearable projector-based gait assistance system and its application for elderly people." Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing. ACM, 2013.

[4]Duong, Thi V., et al. "Activity recognition and abnormality detection with the switching hidden semi-markov model." Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005.

[5]Yang, Jianbo, et al. "Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition." IJCAI. 2015.

[6]Anguita, Davide, et al. "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. " International workshop on ambient assisted living. Springer, Berlin, Heidelberg, 2012.

[7]Anguita, Davide, et al. "A Public Domain Dataset for Human Activity Recognition using Smartphones." ESANN. 2013.

[8]Lara, Oscar D., and Miguel A. Labrador. "A survey on human activity recognition using wearable sensors." IEEE Communications Surveys and Tutorials 15.3 (2013): 1192-1209.

[9]Bechhofer, Sean. "OWL: Web ontology language." Encyclopedia of database systems (2009): 2008-2009.

[10] Ayodele, Taiwo Oladipupo. "Types of machine learning algorithms. " New advances in machine learning (2010).

[11] Anastácio, Ivo, Bruno Martins, and Pável Calado. "Supervised Learning for Linking Named Entities to Knowledge Base Entries." TAC (2011).

[12] Panov, Panče, Larisa Soldatova, and Sašo Džeroski. "Ontology of core data mining entities." Data Mining and Knowledge Discovery 28.5-6 (2014): 1222-1265.

[13] Amores, Jaume. "Multiple instance classification: Review, taxonomy and comparative study." Artificial Intelligence 201 (2013): 81-105.

[14] Stankevich, Evgeny, Ilya Paramonov, and Ivan Timofeev. "Mobile phone sensors in health applications." Proc. 12th Conf. of Open Innovations Association FRUCT and Seminar on e-Tourism. 2012.

[15] Purtov, Konstantin, et al. "Remote photoplethysmography application to the analysis of time-frequency changes of human heart rate variability." Proceedings of the 18th Conference of Open Innovations Association FRUCT. FRUCT Oy, 2016.