A Multi-layer Ontology for Data Processing Techniques

Man Tianxing[1], Nataly Zhukova[1,2], Nguyen Than[1], Alexander Nechaev[4] and Sergey Lebedev[3]
[1]*ITMO University, St. Petersburg, Russia*
[2]*St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia*
[3]*Saint-Petersburg Electrotechnical University, St. Petersburg, Russia*
[4]*Vyatka State University, Kirov, Russia*

Keywords: Data Processing, Machine Learning, Multilayer Structure, Algorithm Selection, Ontology.

Abstract: Currently, data processing technology is applied in various fields. But non-expert researchers are always confused about its diversity and complex processes. Especially due to the instability of real data, the preparation process for extracting information is lengthy. At the same time, different analysis algorithms are based on different mathematical models, so they are suitable for different situations. In the real data processing process, inappropriate data forms and algorithm selections always lead to unsatisfactory results. This paper proposes a multilayer description model of data processing algorithms and implements it based on ontology technology. The model provides a multi-layered structure including data pre-processing, data form conversion, and output model selection so that the user can obtain a complete data processing process from it. The extensibility and interpretability of ontology also provide a huge space for model improvement. The multi-level structure greatly reduces its complexity.

1 INTRODUCTION

A massive amount of domain-specific data which contains useful knowledge are collected, data processing techniques become more and more important in different fields such as marketing, medical, biology, etc (Najafabadi et al., 2015). The researchers are focusing on mining and extracting information from these real-world data sets.

Data collection methods always loosely controlled, resulting in missing values, impossible data combine, noisy value and so on. “Garbage in，Garbage out” means useful information can't be extracted from an incomplete data set (García et al., 2016). The representation and quality of data is the base of a data analysis task. Usually, data preparation and transform take quite an amount of processing time.

The data processing technique is one of today’s most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science (Jordan, M. I. and Mitchell, T. M. 2015).

With the development of related technologies, researchers are always faced with too many choices in dealing with the same data analysis tasks. (Fernández-Delgado et al., 2014) make an incomplete review about available classifiers which already includes 179 classifiers arising from 17 families. And based on the experiments there is no classifier that always has the best performance in every data set. There is no best algorithm, only the most appropriate algorithm. But how to choose the appropriate algorithm is always confusing the non-computer professional researchers.

Anyway, it is a complicated process to generate an object analysis model, which includes many steps. For now, the problem of finding a process to build a model from an initial set is being solved in an ad hoc manner, so it is error-prone and not effective relatively time or efforts

To solve this problem, this article proposes an expandable multilayer conceptual model of data processing techniques. It fixes some stereotypes within the realm of data processing organization, helps to choose among possible alternatives, provides step by step instructions. The model defines possible data formats, data set features, data features, algorithms, user restrictions, data processing workflows. It consists of four layers corresponding to the four forms of data in the data processing process:

- Level 0. Raw layer corresponds to raw data set

Tianxing, M., Zhukova, N., Than, N., Nechaev, A. and Lebedev, S.
A Multi-layer Ontology for Data Processing Techniques.
DOI: 10.5220/0007839606480655
In *Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2019)*, pages 648-655
ISBN: 978-989-758-380-3

- Level 1. The parameter layer corresponds to tidy data set, which is ready to be extracted information by machine learning algorithms.
- Level 2. The indicator layer corresponds to the indicator data which are extracted from raw data set and considered as a new data set.
- Level 3. The processing layer corresponds to the output model.

The authors implement this conceptual model based on ontology technology to build a multilayer ontology for data processing techniques. It describes the processes and suitable situations of data pre-processing techniques, feature extraction algorithms and data processing algorithms. The researchers could get reasonable advice of algorithm selection and complete process description of selected algorithms. The main advantages are as follow:

- This multilayer ontology includes entire process of data processing. Users could find all the information about the data processing techniques in it.
- As an ontology its comprehensibility makes it more friendly to the users and its extensibility make it to be improved in use.
- The multilayer structure split the process of data processing into 4 main steps. This makes the process clearer and such a structure greatly reduces the complexity of the use of the ontology.

This article is organized as follow: the section 2 presents the related work about the existing review of data processing and the techniques which are used in the research; the section 3 describes the construction of the multilayer structure; the section 4 presents the implementation of the Multilayer ontology; the section 5 is the conclusions of this research.

2 RELATED WORK

Data processing is a complex process. Many researchers are committed to providing an excellent taxonomy to help data engineers. Ayodele and T. O. (2010) present a review of the type of machine learning algorithms. Kotsiantis (2007) provide a comprehensive review about Supervised machine learning. Satyanandam and Satyanarayana (2013) describe a taxonomy of ML and data mining for Healthcare Systems. But these reviews just discuss the Theoretical knowledge of data processing techniques. On the other hand, some researchers try to present an understandable introduction about how to choose suitable data processing techniques. Dash and Liu (1997) describe how to select the correct features in classification tasks. Reif et al. (2014) even present an automatic classifier selection model for non-experts. Bernstein et al. (2005) apply ontology technique to build an intelligent assistance for data classification. Anastácio et al. (2011) describe the related knowledge about data mining. Panov et al. (2014) summarize the data mining entities in existing ontologies. These reviews are focus on the part of data analysis. But in fact, in data processing is a complex process, that includes multiple steps starting from data preparation. So the users still don't know how to start with these reviews.

Although some reviews about dealing with the dirty data can be found. Kim et al. (2003) provide a taxonomy of dirty data. Chu et al. (2016, June) describe the methods for data cleaning. García et al. (2015) give a taxonomy of data pre-processing.

Anyway, it takes too much time to check so many literatures to build a data processing process. This article proposes a conceptual model based on the forms of data including the entire data processing process.

Ontology technique is selected to be the method to implement this model. Ontology is a general conceptual model that describes a domain of knowledge (Simons, P., 2000). This model contains the general terms and relationships between the terms in this subject area. It has flexible logical relationships which are suitable for the complex process descriptions in the data processing domain. Its expandability can make the ontology to be expanded with the development of technology so that it will not become obsolete. Its interpretability makes it to be appropriate to the understanding and use of researchers without computer expertise. Keet et al. (2014, July) presented an ontology to describe the knowledge about data mining. Rodríguez-García et al. (2016) presented a semantically boosted platform for assisting layman users in extracting a relevant subdataset from all the data and selecting the data analysis techniques.

Multi-layer concept is effective for the data conversion process (Osipov et al., 2017). The concept of multi-layer ontology is also used to implement synthesized models. Pai et al. (2017) create a multi-layer ontology-based information fusion for situation awareness. CARVALHO, V. (2016) presents the main method to build multi-layer ontology conceptual model.

So this article present a multi-layer conceptual model of data processing techniques. The forms of data are the basis for splitting the process. A multi-layer ontology is created as the implement of this conceptual model.

3 MODEL STRUCTURE

The process of data processing, which extracts information from the raw data, is complex. In order to provide a clear structure, the authors designed a multi-layer conceptual model based on the data forms in data processing process. The essence of data processing is to constantly transform the data set until it becomes understandable knowledge. These conversion operations are data processing techniques. In this model data sets in different layers are converted to each other with the operation of data processing techniques. However, users only need to consider the dataset characteristics and the available algorithms in the current layer. Such a structure greatly reduces the complexity of synthesizing the entire data processing process.

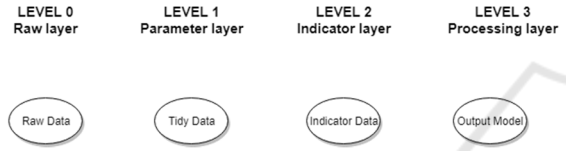

Figure 1: The multi-layer structure of the conceptual model.

The multilayer structure is shown in Figure 1. The data changes are divided into four levels. Pre-processing techniques, feature extraction methods and machine learning algorithm are applied as data conversion operations.

Table 1: Main data defects and their corresponding pre-processing technique.

Data defect	Pre-processing technique
Missing value	Delete, Ignore, Imputation etc.
Redundancy	Reduction
Noisy value	Ensemble Filter, Iterative-Partitioning Filter etc.

- Level 0: The raw layer is used to describe the raw data set in real life due to the poor management of data collection methods, many anomalies are recorded on real-world data set. These serious data defects make machine learning algorithms not directly applicable. So this level represents the defects of raw data set and the corresponding pre-processing techniques which are shown in table 1.
  - Level 1： The output of data pre-processing is actually the final train set for machine learning algorithms. The tidy data means that its form is ready to be analysed. But the characteristics of data set are the important factor for choosing suitable machine learning algorithm. So in this level the data is marked with their characteristics which are shown in table 2.

Table 2: Main characteristics of tidy data set.

Category	Data characteristic
Sample size	Great, Medium, Small
Attribute size	Great, Medium, Small
Relevance	Irrelevant, Related
Type	Time series etc.

- Level 2: As mentioned earlier, the initial tidy data is only formally ready to be analysed. Each of its parameters has relatively complete and reasonable data. But the original parameters do not vividly express the needs of some users. The data needs to be described with appropriate features to represent the corresponding properties. This situation is especially common in time series data analysis. The data is subjected to feature extraction and new parameters appear as a new data set. The choices of the extracted features depend on the data properties that the user desires. Table 3 presents different features which can represent different properties.

Table 3: Main data property and their measures.

Data property	Data feature
consistency	intervals
dispersion	Quartile Deviation, Mean Deviation, skewness, kurtosis etc.
central tendency	Arithmetic mean, Geometric mean, Harmonic mean, median, mode, quantiles etc.

- Level 3: The output model is the final form of data and the main goal of data processing. Neither the classification model nor the cluster model represents the knowledge extracted from the data set. So this layer represents the user's needs most directly. In the multi-layer conceptual model this layer is described as the characteristics of output model which is shown in table 4.

Table 4: Main characteristics of output model.

Category	Model characteristic
Type of task	classification, cluster, prediction etc.
number of classes	non-class, two-classes, multi-classes
Interpretability	Interpretable, Inexplicable

These 4 forms of data exist in the entire process of data processing and the process actually is the mutual conversion between data forms. As the Figure 2 shown raw data is prepared by pre-processing techniques step by step until it becomes tidy enough. Unless the user has special needs, tidy data can be analysed by machine learning algorithms to generate an output model. However, sometimes tidy data must be converted by feature extraction methods to generate a new data set whose parameters are the features of original data. And this new data set is a new raw data. It is very possible that this new data set loses the tidy form, so it should be pre-processed again to be ready for analysis.

The data at each level has unique characteristics and problems to be solved. Data processing techniques should be applied with the right sequence. Without consideration of the states of data, some repetitive operations will be applied. This situation increases the complexity of data processing work. Such as principal components analysis should be used after representation, because the representation methods can reconstruct the structure and content of the data set, which offset the effect of principal components analysis.

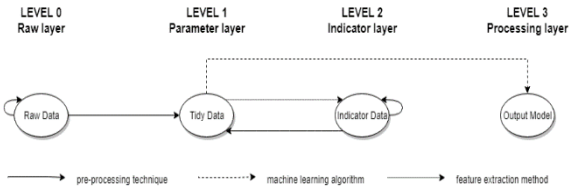

Figure 2: The conversion in the multi-layer conceptual model.

However, the multilevel structure makes users just consider the corresponding problems in each level. The irrelevant conditions couldn't be seen. With the transform among the data levels, a reasonable operation process is generated.

4 IMPLEMENT WITH ONTOLOGY

Ontology is a commonly used technique for describing domain-specific knowledge. Its extensibility and comprehensibility are suitable for the construction of the conceptual model in this paper.

The authors build the multi-layer ontology of data processing techniques on protege-5.5.0 using the owl language (McGuinness and Van Harmelen, 2004). It is inferred by Hermit 1.3.8.413 reasoner. There are no highlight classes appeared and the defined individuals are classified into the correct classes. It means this ontology satisfies consistency and sufficiency.

4.1 Basic Structure of Multi-layer Ontology

Figure 3 shows the basic structure of the multi-layer ontology. It is obvious that description of different data forms only connects to the appropriate algorithm for themselves so that the information is split into different levels.

The multi-layer structure is constructed based on the relationships in ontology. All the data processing techniques are linked the characteristics of data in different level. The relationships are the object property "suitableFor". When user deal with a data set, they only can see all the available solution based on the outward links of the characteristics of this data set. And if the data set is converted to another form. Users can find the suitable solutions based on the outward links of the characteristics of the new data set.

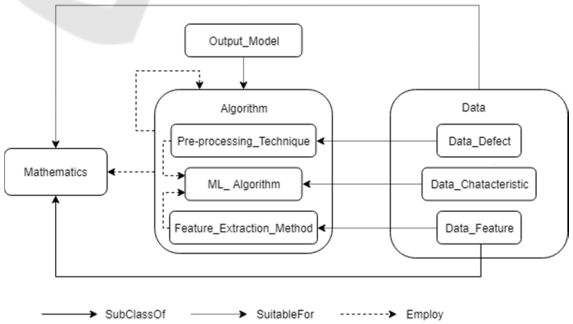

Figure 3: The basic structure of the multi-layer ontology.

4.2 Classes in Multi-layer Ontology

Classes are the core content of the ontology. In order to provide users with clear data processing knowledge, this multi-layer ontology creates several main classes:

- Algorithm: This is the main part of the ontology. Pre-processing techniques, feature extraction methods, and machine learning algorithms are all subclasses of this class. Hundreds of data processing techniques are described in this class. It also provides a variety of taxonomies to classify these algorithms so that users can understand these algorithms from different angles.
- Mathematics: This class describes the mathematical knowledge involved in data processing. Because the performance differences of the algorithm depend on the mathematical foundation such as measure, models etc. It is an important basis for algorithm selection.
- Data: This class contains various information of data such as data defect, data characteristic, data property etc. It describes the characteristics of the data in each level.
- Output model: This class describes the users' requirement. It uses the characteristics to represent tasks.

These classes describe the knowledge about data processing. This is a comprehensive and extendable review so that users can achieve enough information from it.

4.3 Properties in Multi-layer Ontology

The definition of property is the greatest difference between ontology and taxonomy. In ontology more properties are defined to present more complicate relations. In this multi-layer ontology both object property and data property are used.

4.3.1 Object Property

The main object properties are as follow:

- "subclassOf" is a typical relation in ontology and taxonomy. It makes the hierarchy of algorithms and data information clearer. And many taxonomies are integrated into this ontology based on this relation.
- "employ" is used from class algorithm to class mathematics to explain the algorithm process. Sometimes the data sets are suitable for some special measures or functions. So it connects algorithms and mathematics to lead to find the suitable algorithm with suitable measure.
- "suitableFor" is used to provide advice about algorithm selection. The relations "suitableFor" are from theory base and some previous experiment results.

4.3.2 Data Property

Data property is just used to describe the value and range of machine learning algorithms. This part is very useful for non-expert researchers. The table 5 is an example about data property.

Table 5: The use of data property “hasSize” in ontology.

Category	Range of Class Value	Entity in Ontology
Sample size	hasSize some xsd:integer[< 100]	SmallTrainDataset
	hasSize some xsd:integer[>= 100, <=500]	SmallTrainDataset
	hasSize some xsd:integer[> 500]	LargeTrainDataset
Attribute size	hasSize some xsd:integer[< 300]	ShortDataset
	hasSize some xsd:integer[>= 300, <=700]	MediumDataset
	hasSize some xsd:integer[> 700]	LongDataset
No. of classes	hasSize some xsd:integer[< 10]	FewClassDataset
	hasSize some xsd:integer[>= 10, <=30]	MediumClassDataset
	hasSize some xsd:integer[> 30]	ManyClassDataset

4.4 Workflow of Multi-layer Ontology

Based on the object property defined in the ontology, users can find the right solution at each level. Figure 4 is the workflow based on this multi-layer ontology.

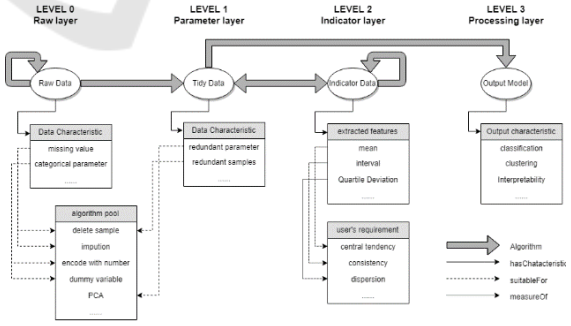

Figure 4: The workflow of the multi-layer ontology.

The data is converted by the class “Algorithm”. Since the data characteristics link their available solutions with the object property “suitaleFor” in the ontology, users don’t need check the whole algorithm pool. And in each level users just choose the suitable

solution based on the data characteristics and their requirements.

5 APPLICATION

This multi-layer ontology is quite meaningful for researchers who are not computer professionals. The authors used this ontology to process real-world data in various fields and achieved good results (Tianxing and Zhukova, 2018). This article presents an application to estimation of the complex object states.

In this experiment, the data set is the results of measurements of technical objects parameters. The user wants to make system analyses of the objects states. Each object has some characteristics and users want to find some deep information inside.

The original data set has 534 samples for 89 objects. And each object has up to 6 samples corresponding to 6 state points. Each sample has 21 attributes corresponding to 21 parameters.

The original objects have already been classified to 3 group: Group 1 and group 2 are two different categories and group 3 is for the objects whose data is not enough. In this experiment, we only try to find the difference of the consistency of changes in the parameters.

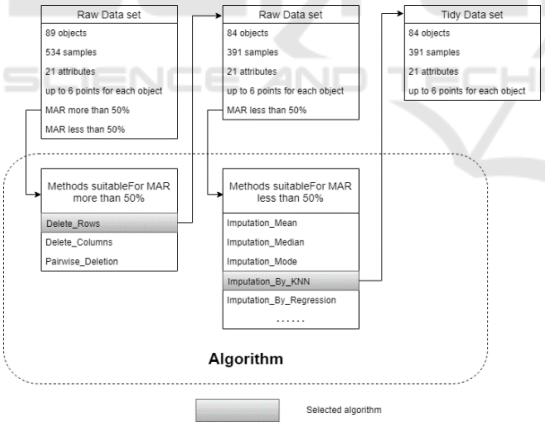

Figure 5: Data conversion from level 0 to level 1.

Users are often confused in the face of a messy data set and abstract task requirements. They don't know where to start. Choosing the right set of algorithms in hundreds of data processing algorithms to generate the correct process is a complex task. The multi-layer structure provided by multi-layer ontology simplifies this process. When users enter current data characteristic, they only get a collection of available solutions. For example, the raw data set has several defects. But each time the user enters a defect, they get a collection of suitable solutions. After solving all the defects with the provided solutions， the raw data set has no defects, then it is converted to a tidy data set. Then user will just get available solutions in the algorithms which are available for tidy data. Therefore, the complexity of data processing synthesis is greatly reduced. The detailed operation flow is as follows:

First step is obviously to prepare the data which is shown in Figure 5. Missing value is the problem of this data set. Two solutions were chosen to solve two different missing situations to get the tidy data. The level changes from level 0 to level 1.

In order to verify the difference in the data of different groups, Authors decided to test the consistency of the changes between points in different states. But it is difficult to express consistency from the original parameters. So as shown in the figure 6, interval is selected as the new parameter. The level changes from level 1 to level 2.

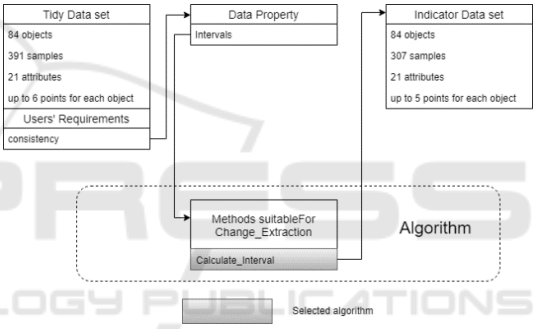

Figure 6: Data conversion from level 1 to level 2.

However, the main idea is clustering the intervals value to check consistency. The values in indicator data have different ranges so that the clustering algorithms couldn't achieve good results. So normalization operation is applied on the indicator data to convert it from level 2 to level 1 so that it is ready for a clustering algorithm as the Figure 7 shown.

At last a clustering algorithm "EM" is selected based on the characteristics of data and output model as the Figure 8 shown.

And the result of clustering intervals between 1st point and 2nd point in group 1 and group 2 with EM algorithm is shown in table 6.

The intervals in group 1 are clustered into 1 cluster. That means the data is compact and respectively changes of parameters are consistent in group 1. On the other hand, the distribution in group 2 is blurred. The Figure 9 shows the example of the probability distribution of an indicator. This conclusion is a significant result of the technical object analysis.

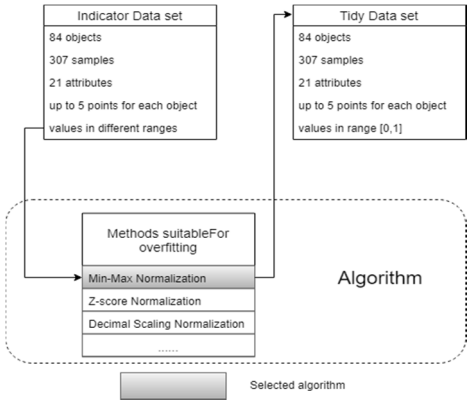

Figure 7: Data conversion from level 2 to level 1.

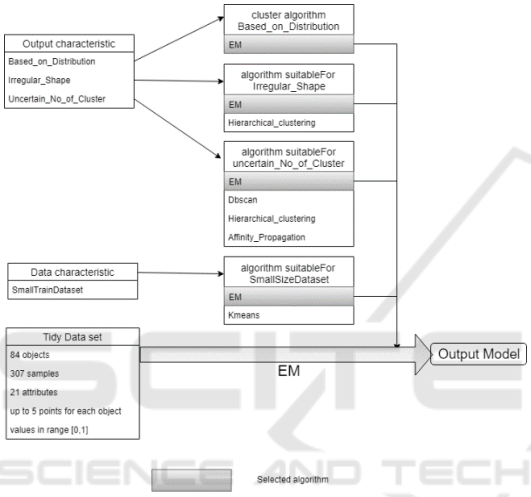

Figure 8: Data conversion from level 1 to level 3 (EM-Expectation-Maximization algorithm).

Table 6: The results of clustering intervals between 1st point and 2nd point in group 1 and group 2 with EM algorithm.

dataset	No. of samples	No. of cluster	likelihood
group1_diff1 2_norm	29	1 {29}	89.23071
group2_diff1 2_norm	31	2 {8,23}	92.30114

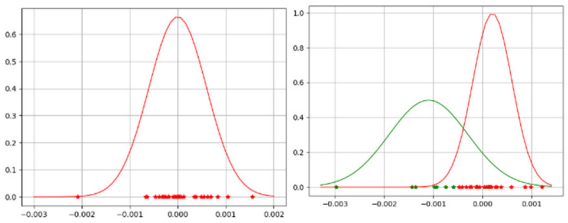

Figure 9: The probability distribution of 1st intervals of group 1 (1 cluster) above and group 2 (2 clusters: 1st cluster in green colour and 2nd cluster in red colour) below.

6 CONCLUSIONS

For the complex structure of data processing, a clear conceptual model can effectively save time and computational resources in process synthesis. This paper proposes a conceptual model of data processing technology based on the form of data transformation. It includes the entire data processing process including data pre-processing, feature extraction and data analysis. The multi-layer structure divides these methods into four levels: raw layer, parameter layer, indicator layer and processing layer. Such a structure reduces the complexity of the process synthesis. The authors implemented this conceptual model using ontology technology. Based on the advantages of the ontology, this multi-layer ontology is also extensible and understandable. The experiments prove that the multi-layer ontology provides sufficient information about the synthesis of real-world data processing process and it is significant to the non-expert users.

REFERENCES

Kim, W., Choi, B. J., Hong, E. K., Kim, S. K., & Lee, D. (2003). A taxonomy of dirty data. *Data mining and knowledge discovery,* 7(1), 81-99.

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016, June). Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data* (pp. 2201-2206). ACM.

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining* (pp. 59-139). New York: Springer.

Pai, F. P., Yang, L. J., & Chung, Y. C. (2017). Multi-layer ontology-based information fusion for situation awareness. *Applied Intelligence*, *46*(2), 285-307.

CARVALHO, V. (2016). Foundations for Multi-level Ontology-based Conceptual Modeling.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1.

García, S., Ramírez-Gallego, S., Luengo, J., Benitez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*, *1*(1), 9.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. *The Journal of Machine Learning Research*, *15*(1), 3133-3181.

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, *1*(3), 131-156.

Reif, M., Shafait, F., Goldstein, M., Breuel, T., & Dengel, A. (2014). Automatic classifier selection for non-experts. *Pattern Analysis and Applications*, *17*(1), 83-96.

Bernstein, A., Provost, F., & Hill, S. (2005). Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification. *IEEE Transactions on knowledge and data engineering*, *17*(4), 503-518.

Anastácio, I., Martins, B., & Calado, P. (2011). Supervised Learning for Linking Named Entities to Knowledge Base Entries. *TAC*.

Panov, P., Soldatova, L., & Džeroski, S. (2014). Ontology of core data mining entities. *Data Mining and Knowledge Discovery*, *28*(5-6), 1222-1265.

CARVALHO, V. (2016). Foundations for Multi-level Ontology-based Conceptual Modeling.

Osipov, V., Vodyaho, A., & Zhukova, N. (2017). About one approach to multilevel behavioral program synthesis for television devices. *International journal of computers and communications*, *11*, 17-25.

Tianxing, M., & Zhukova, N. (2018). An Ontology of Machine Learning Algorithms for Human Activity Data Processing. *learning*, *10*, 12.

Synthesis of integral models of system dynamics of an acid-base state of patients at operative measures, http://www.actascientific.com/ASMS-3-2.php

Simons, P. (2000). Parts Study in Ontology: A Study in Ontology.

McGuinness, D. L., & Van Harmelen, F. (2004). OWL web ontology language overview. *W3C recommendation*, *10*(10), 2004.

Keet, C. M., d'Amato, C., Khan, Z. C., & Lawrynowicz, A. (2014, July). Exploring Reasoning with the DMOP Ontology. In *ORE* (pp. 64-70).

Rodríguez-García, M. Á., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., García-Sánchez, F., & Valencia-García, R. (2016). Ontology-Based Platform for Conceptual Guided Dataset Analysis. In *Distributed Computing and Artificial Intelligence, 13th International Conference* (pp. 155-163). Springer, Cham.