A Knowledge-Based Computational Environment for Real-World Data Processing

Man Tianxing[1], Nataly Zhukova[1,2], Vasily Meltsov[3], and Yulia Shichkina[4(✉)]

[1] ITMO University, St. Petersburg, Russia
mantx626@gmail.com, nazhukova@mail.ru
[2] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia
[3] Vyatka State University, Kirov, Russia
meltsov69@mail.ru
[4] St. Petersburg State Electrotechnical University "LETI", St. Petersburg, Russia
strange.y@mail.ru

Abstract. The data contains knowledge. Researchers working in various fields are struggling to extract information from specific data sets. But data processing is a complex process. Real-world data must be converted multiple times to become useful knowledge. This process is very unfriendly to non-computer professional researchers. The lack of knowledge related to data processing makes their work unsatisfactory. The main purpose of this paper is to build a computational environment based on ontology technology, which provides selection and description of data processing algorithms to help the users extract information from the real-world data. This article provides a real case to indicate the rationality of this computational environment. The authors extracted significant conclusions from a data set of the acid-base state at patients.

Keywords: Data processing · Ontology · Computational environment

1 Instruction

With the rapid development of information technology, researchers in various fields are paying more and more attention to real-world data contained the knowledge. Extracting significant knowledge from the data generated in daily life is a fantasy work. But data processing is a quite complex process so that many non-computer professional researchers are confused about the construction of the entire data processing process. In fact, most researchers who are trying to extract information from data are not working in the Internet industry, such as market analysis, medical, financial investment and so on.

The main purpose of this paper is to create a knowledge-based computational environment (KBCE) for real-world data processing based on ontology technology. Based on the KBCE users can select appropriate data processing algorithms and build the entire data processing process according to the characteristics of the data set and the requirements of the task step by step.

© Springer Nature Switzerland AG 2019
S. Misra et al. (Eds.): ICCSA 2019, LNCS 11619, pp. 257–269, 2019.
https://doi.org/10.1007/978-3-030-24289-3_20

KBCE is an ontology that contains the knowledge about data processing and its logical relationships. The authors create the classes of algorithms, mathematics, input features, output features, and process, and define object property that expresses their inner relationships and data property that describes the specific parameter settings of the classes. Because it is based on ontology technology, KBCE expresses more logical relationships more explicitly than traditional taxonomies. The advantages of KBCE are as follows:

- KBCE is different from other reviews about data processing. It describes the entire data processing process including data preprocessing, classification, clustering, and even evaluation modules to provide a complete workflow for the user. A large number of facts prove that in data mining systems, data preprocessing accounts for 60% to 80% of the total workload [1]. A complete description of data preprocessing is significant to real-world data processing.
- KBCE is built based on the ontology technology, so it can be easily expanded according to the requirements of users. Even users can customize the specialized computing environment by linking knowledge in the field.
- In addition to the description of the algorithmic process, KBCE also summarizes the performance and application of data processing algorithms based on a large number of experiments and articles. The user could input the characteristics of the data set and the task requirements to get the appropriate solution. With KBCE users can make decisions based on the true performance of the algorithm rather than their intuition.
- The expression of ontology is user-friendly. Although users may have no experience in the computer industry, as an ontology KBCE presents clear logical definition and relationships like a mind may so that they can understand their selection and the entire data processing process.

The structure of this paper is as follows: Sect. 2 introduces background knowledge and related work. Section 3 presents the structure and the main components of KBCE. Section 4 describes the workflow of KBCE. In Sect. 5 the authors apply KBCE on a real-world data set about the acid-base state (ABS) of the patients to achieve some medical conclusions to indicate the rationality of KBCE. In Sect. 6 the authors present the conclusions of KBCE and future work.

2 Background

2.1 Data Processing

With the rapid development of the information industry, people have accumulated more and more data. The real-world data contains a lot of important information. So how to analyze data at a higher level to make it significant becomes essential. Data processing has become a very active frontier in computer science. In fact, most of the researchers, who are trying to extract information from data, work in non-computer fields, such as market analysis, medical, climate analysis [2–4] and so on.

Many studies are applying the methods and models on ideal data rather than the intricate real-world data sets. In general, it is inevitable that there are redundant data, missing data, uncertain data and inconsistent data [5–7]. According to the principle of "garbage in, garbage out" [8], inappropriate data formats can result in expensive operating costs and long response time and affect the correctness of the patterns extracted from the data set and the accuracy of the output models. So the real-world data processing process is a complex process form multiple conversions and the final information extraction as the Fig. 1 shows.

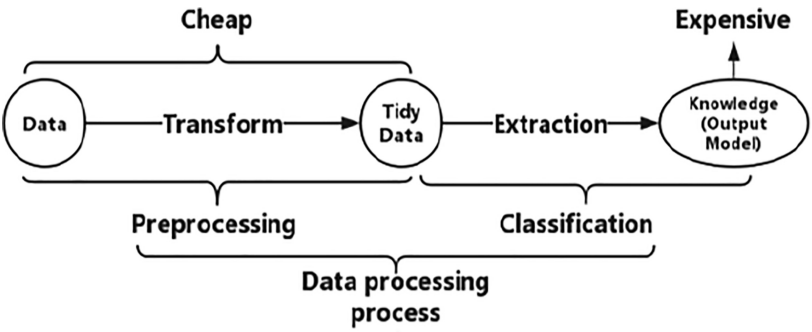

Fig. 1. General data processing process

2.2 Reviews of Recommendation Systems of Data Processing

Fernández-Delgado did a series of interesting experiments about applying more than one hundred classification algorithms on different data sets to present "There is no best classifier, only the most suitable classifier" [9]. This situation is also existing in data processing. The size, quality and type of data sets and users' requirements all affect the performance of the algorithm. The algorithm which is suitable for all the situations doesn't exist.

Researchers have provided some solutions to provide data processing users with advice on algorithmic choices. Scikit-learn is the most commonly used machine learning (ML) python package for python users. Its developers also provide a classic ML algorithm selection scheme [10]. Some computer science researchers have proposed various automatic selection methods for supervised machine learning problem [11]. The practical application effects of clustering algorithms are also summarized [12]. Some studies focus on the processing of specific data forms. Bagnall [13] and Fawaz [14] present great reviews of time series classification algorithms. Although these reviews provide excellent content. But too much professional knowledge is confusing the researchers who don't work in computer science. The purpose of this article is to present a user-friendly and complete computational environment of data processing.

2.3 Ontology Technology

Ontology is a formal specification for communication and knowledge sharing between different subjects in the domain by defining the relationships among the concepts that are commonly recognized in the field by describing and capturing domain knowledge [15]. It is an explicit, formal, shared conceptualization. So it can be regarded as a formal knowledge, even a knowledge of management knowledge which can satisfy the academic research community to share the expression of knowledge. The conceptual model is ideal for building recommendation and interpretation systems. Some researchers have tried to apply the ontology technology to the algorithm recommendation system [16], but such systems are only focused on specific purposes so that they are not extensive. The purpose of this paper is to create an ontology about data processing as a knowledge interface in the field of data processing and an understandable computing environment for non-computer professionals.

3 Construction of KBCE

3.1 Basic Structure

The KBCE is an ontology which is written with OWL language and available at https://github.com/529492252/KBCE-ontology.git. It is edited in Protégé-5.5.0 and was checked by HermiT 1.3.8.413 reasoner to make sure it is consistent.

KBCE has two main parts: the creation of class and the definition of property which are shown in Fig. 2. They are used to describe the logic relationships about the process and performance of data processing. The basic structure is in Fig. 3.

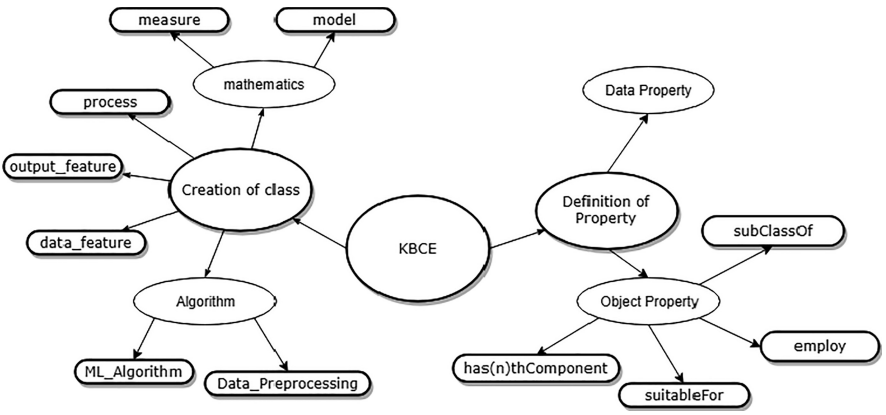

Fig. 2. The main components of KBCE

In KBCE most of the classes are concrete operations, but the authors also define some abstract concepts to make the logic clearer. And object property is used to describe the logic as Fig. 3 shown. In ontology the other kind property "data property" is still important to provide the parameter settings of algorithms.

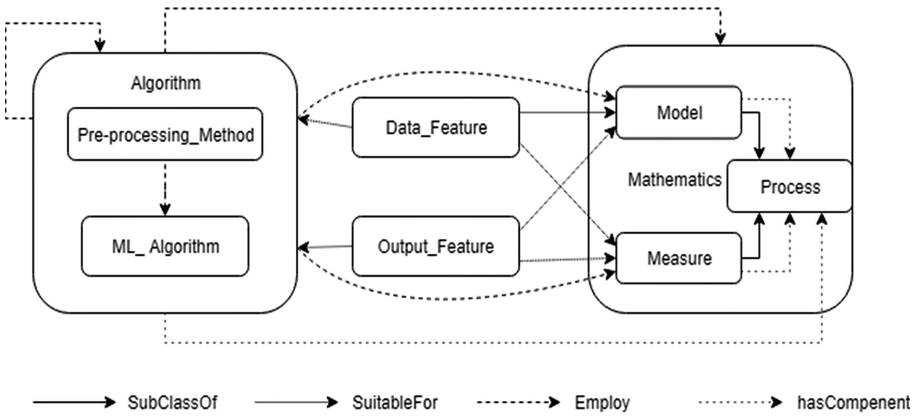

Fig. 3. The basic structure of KBCE

3.2 Creation of Classes

Knowledge is the foundation of KBCE. It is difficult to sort out the knowledge related to data processing because there are too many different angles of taxonomies. In the conceptual model of ontology, class means thing with certain attributes. Such a statement is precisely suitable for describing complex data processing knowledge. The authors mainly create such classes:

- Algorithm - algorithms for data processing that can be common ML algorithms and special methods oriented on data preprocessing. Figure 4 present part of the class "Algorithm".
- Data_Feature - characteristics of input datasets, including size, length etc.
- Output_feature - features of the output such as the number of classes of classification task. And the users' requirements are classified to this group like speed, accuracy, interpretability etc.



Fig. 4. Part of class "Algorithm" (main dimensionality reduction methods are shown)

- Mathematics - mathematics base for the algorithms. This is a big group because it contains most of operations of algorithms. Such as Measure – measures between data in corresponding similarity functions; Algorithm model – basic algorithms models.
- Process - This is an abstract concept classes which the authors create for describing the process of algorithms. Authors name all the operations in each algorithm and specify them in this group. When users try to employ an algorithm without any information about it, they look through the content about this algorithm in "process" to get the process.

3.3 Definition of Property

Object Property. Ontology can define more properties to present relations. This is the greatest difference from taxonomy which only has a relationship "has-a". In KBCE some main object properties are defined:

- subclassOf - basic property to link father class to subclass.
- employ - links the algorithm and their core content. Such as measures and models that can affect the performance of algorithms.
- has(*n*)thComponent - Links the algorithms and their components to describe the processes of the algorithms. The authors define the steps of the algorithm by assigning a value to *n* present the sequence of the operations.
- suitableFor - links input data characteristics and users' requirements to the algorithms, defining suitable algorithms for processing data with known characteristics.

Data Property. Since KBCE is oriented to non-computer professional researchers, the parameters setting is usually a problem for them. Data property is used to define the value or range of the operations. And when users want to input a data characteristic, the data property can guide users to translate it into a class in KBCE.
As the Table 1 shown the data characteristics are defined by data property "hasSize", if user input a value of the size of train data set, Length or No. of classes, the reasoner of KBCE can translate it into a class in ontology automatically.

Table 1. The data characteristics are defined by data property "hasSize"

Category	Range of Class Value	Class in KBCE
Train size	hasSize some xsd:integer[< 100]	SmallTrainDataset
	hasSize some xsd:integer[>= 100, <=500]	SmallTrainDataset
	hasSize some xsd:integer[> 500]	LargeTrainDataset
Length	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 Generation of Data Processing Process Based on KBCE

KBCE is primarily for non-computer professionals. Its purpose is to clearly present the logical structure of data processing knowledge so that users can perform fewer steps to get better results. The generation of the KBCE data processing process is to sequentially input the task conditions to obtain a suitable set of algorithms. Finally choose the most suitable algorithm. Specific steps are as follows:

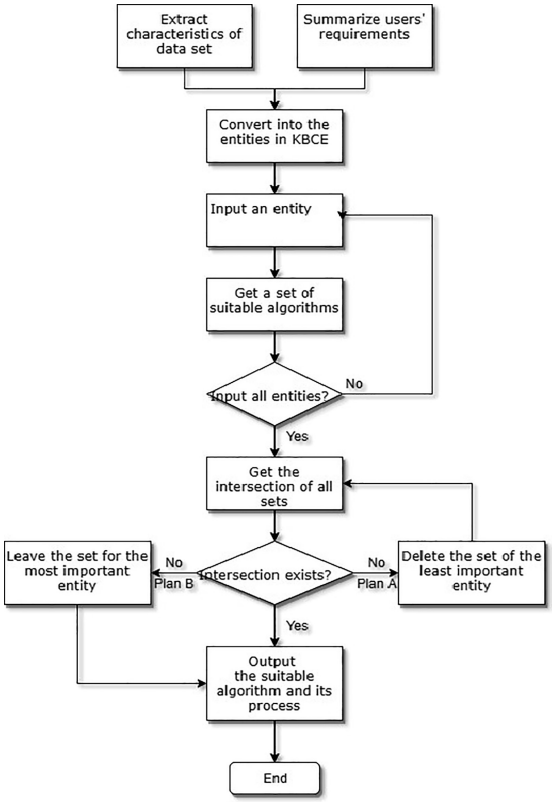

Fig. 5. The workflow of KBCE

1. Acquire user requirements and extract the characteristics of the data set.
2. Enter the task conditions and convert them to entries in KBCE.
3. Get the appropriate set of algorithms in turn.
4. Extract the intersection of the set of suitable algorithms.
5. If the intersection does not exist, Plan A deletes the least important entry and then performs step 3; Plan B is directly selected in all current suitable algorithms (shown in Fig. 5).
6. Output the selected algorithm.
7. (Optional) Describe the selected algorithm based on the outward links.

5 Case Study

To indicate the feasibility of KBCE, the authors analysis a real-world data set about acid-base state (ABS) of patients to extract information based on KBCE. The dynamics of the ABS in cavernous sinus (CS) was studied for 89 patients with cardiac surgical pathology during the postoperative period in the operating room and in the cardio-resuscitation unit. Doctors want to make a systematic analysis on it.

The original data set has 534 samples for 89 patients that each of them has up to 6 samples corresponding to 6 state points. Each sample has 42 attributes which come from 21 parameters with a pair of attributes. The parameters are shown in Table 2. Paired attributes are one-to-one correspondence. So in this experiment 21 attributes are selected. The patients are classified to 3 group: patients in group 1 have healthy physical state and patients in group 2 usually can recover soon when the parameters change, and group 3 is for the patients whose data is not enough.

Table 2. Parameters of ABS

Parameter name	Description of the parameter	Parameter name	Description of the parameter
pH	Acidity	Na+	Sodium ion concentration
pO2	Oxygen partial pressure	Ca ++	Calcium ion concentration
pCO2	Carbon dioxide partial pressure	Cl-	Chlorine Ion concentration
ABE	Excess base	Glu	Glucose concentration
SBE	Lack of reason	Lac	Lactate content
cHCO3	Plasma bicarbonate	p50	Hemoglobin affinity for oxygen
cHCO3-st	Bicarbonate (alkali)	mOsm	Blood osmolarity
sO2	Oxygen boost	pH(T)	Acidity corrected for temperature
ctHb	Reference hemoglobin level	pO2(T)	Partial oxygen pressure adjusted for temperature
Htc	Hematocrit	pCO2(T)	Carbon dioxide partial pressure adjusted for temperature
K+	Potassium ion concentration		

The doctor hopes to extract intrinsic characteristics of parameters change for the patients in different groups. It should be important to distinguish patients in group 1 and group 2. General classification algorithms can classify the patients precisely. So this experiment is used to figure out the coherent consistency of the parameters in this data set.

The average dynamics of ABS is given in Fig. 6. Authors hope to extract more information from this data set. With the intuition they applied some cluster algorithms on the data set and compared with the existing labels. But the results are unsatisfactory. So they tried to analysis this data set based on KBCE for extracting deep knowledge.

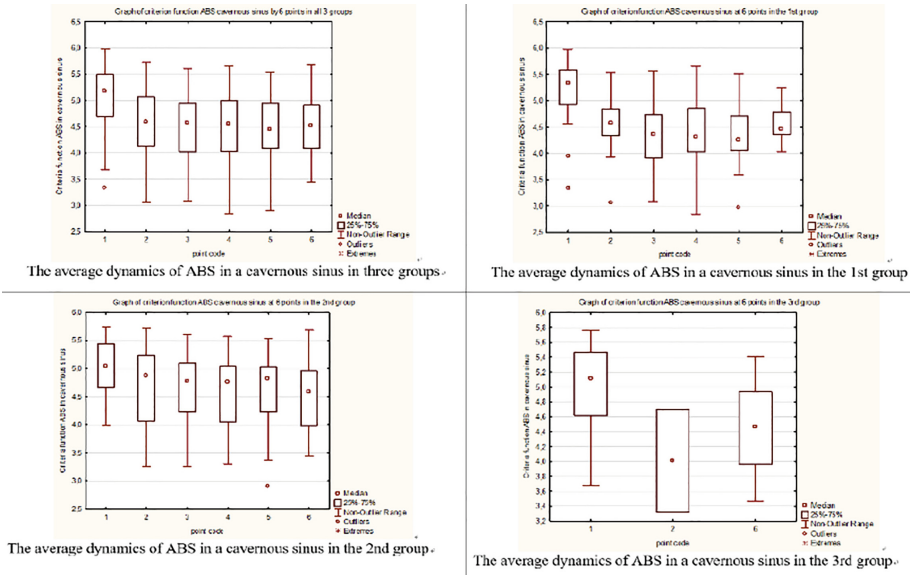

Fig. 6. The average dynamics of ABS in a cavernous sinus in all groups and in each group [17]

The original data set is not tidy enough, several preprocessing operations are provided by KBCE as Figs. 7 and 8 shown. There are too many missing values, which is missing at random (MAR), in the original data set. KBCE classifies them into two groups: MAR more than 50% means more than 50% values are missing in each row or column or both of them; MAR less than 50% means less than 50% values are missing in each row or column. KBCE suggests to "Delete_Rows" for MAR more than 50%

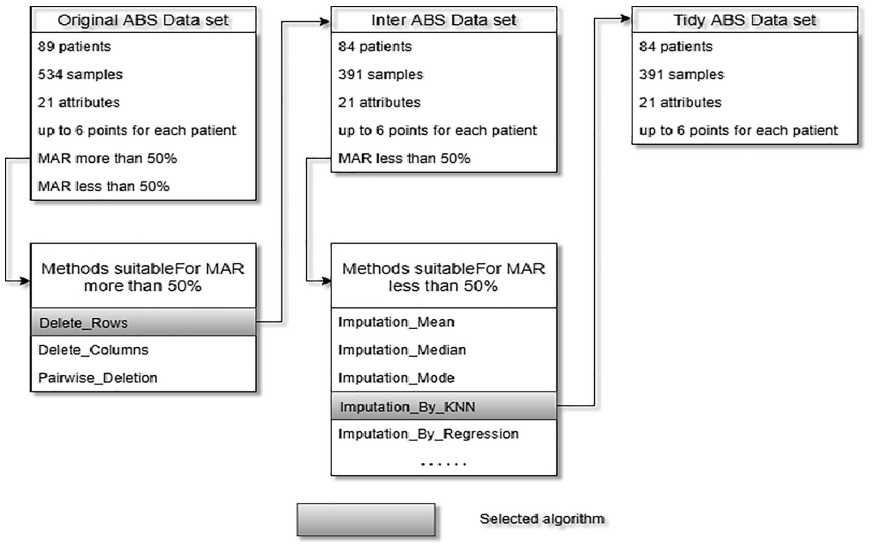

Fig. 7. Preprocessing based on data set characteristics

because some samples in data set lose too many values. And KBCE suggests using imputation methods to deal with MAR less than 50%. According to the data characteristic "interdependent" the system selects "Imputation_By_KNN" (KNN - k-Nearest Neighbor Algorithm).

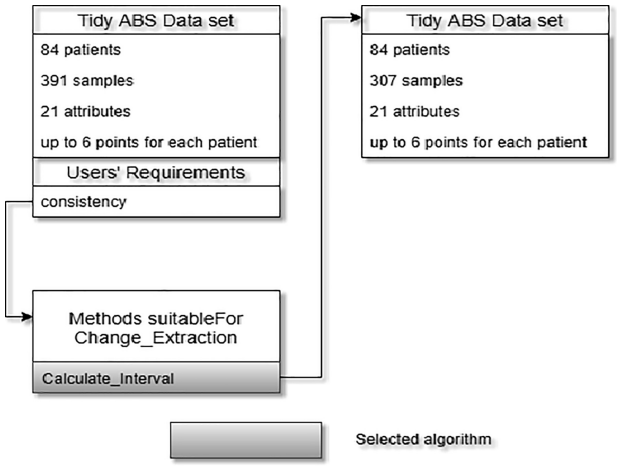

Fig. 8. Preprocessing based on user's requirements

As the Fig. 8 shown consistency of the parameters is user's requirement, and KBCE converts the requirement to the analysis of the variations between different points in different groups. KBCE suggests that to analysis the intervals is the best method to present the changes in data sets.

After preprocessing the data set is tidy enough for clustering. The system selects the most suitable cluster algorithm based on the information in KBCE which is presented in Fig. 9. For checking the consistency of data set KBCE must convert consistency into some entities in system. Soft cluster is significant for the flexible clustering, so "Based_on_Distribution" is selected. And the data represents different parameters, so the clusters should be "Irregular_Shape". The number of clusters shouldn't be set, because it means the consistency of the parameters. At last they consider the size of data set is small. So "SmallSizeDataset" is added for a reference. Then they input the entities in KBCE and get the suitable algorithms for each entity. Usually the intersection of all the sets of algorithms which are suitable for each entity is the output of KBCE.

In this experiment the EM algorithm satisfies all the entities which are converted from user's requirement, especially it is based on distribution, so it could show the consistency among the data set. And Based on the outward links of EM it is split into steps: Normalization and cluster. So the entire process of ABS data set processing is presented:

Delete_Rows -> Imputation_By_KNN -> Delete_Useless_Samples -> Calculate_Intervals -> Normalization -> EM

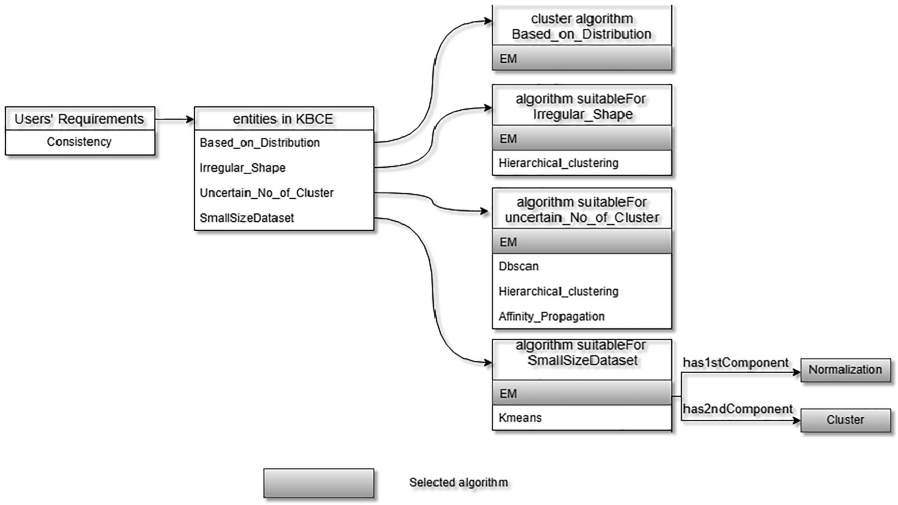

Fig. 9. Select the most suitable cluster algorithm in KBCE (EM- Expectation Maximization)

Based on EM algorithm the authors consider the changes among all the points in different groups. They present interesting results. The comparison of the intervals between 1st point and 2nd point in group 1 and group 2 is a typical example which is presented in Table 3.

Table 3. 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_diff12_norm	29	1 {29}	89.23071
group2_diff12_norm	31	2 {8,23}	92.30114

According to the table content, 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. This conclusion is significant for doctor's work. In Figs. 10 and 11 the probability distribution of the 1st and 33th intervals in group 1 and group are separately shown.

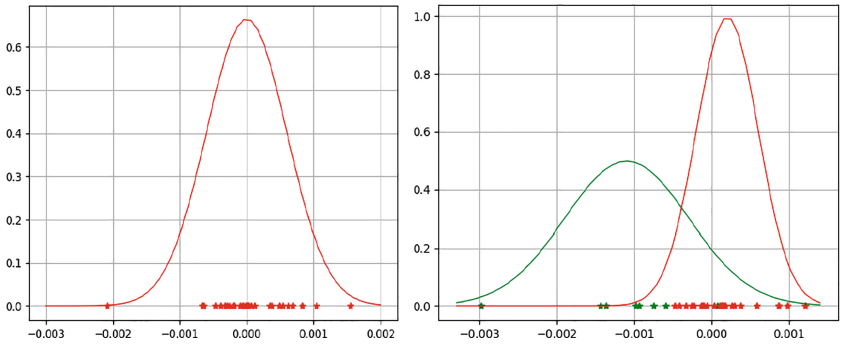

Fig. 10. The probability distribution of 1st intervals of group 1 (1 cluster) and group 2 (2 clusters: 1st cluster in green color and 2nd cluster in red color) (Color figure online)

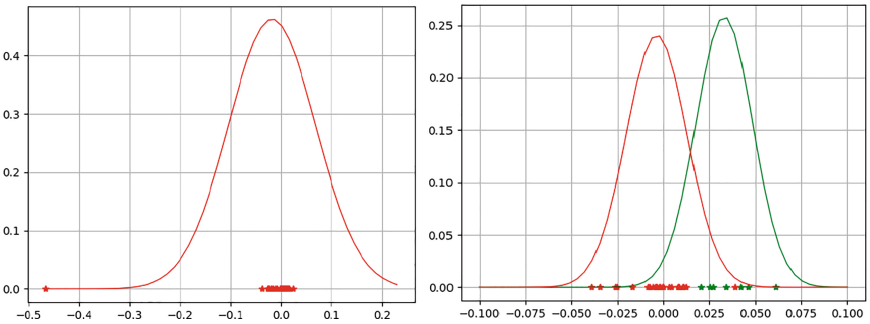

Fig. 11. The probability distribution of 33th intervals of group 1 (1 cluster) and group 2 (2 clusters: 1st cluster in green color and 2nd cluster in red color) (Color figure online)

6 Conclusion

This paper presents a knowledge-based computational environment for real-world data processing. It is built based on ontology so that it is flexible and understandable for non-computer professional researchers. Users can get reasonable advice about selecting the suitable data processing algorithm and understand these algorithms very easily according to the logical description in KBCE. The case indicates that KBCE usually provides a complete and reasonable data processing solution.

The research was funded by RFBR and CITMA according to the research project no. 18-57-34001.

References

1. Romano, D.: Data mining leading edge: insurance & banking. In: Proceedings of Knowledge Discovery and Data Mining (1997)
2. Stankova, E.N., Balakshiy, A.V., Petrov, D.A., Shorov, A.V., Korkhov, V.V.: Using technologies of OLAP and machine learning for validation of the numerical models of convective clouds. In: Gervasi, O., et al. (eds.) ICCSA 2016. LNCS, vol. 9788, pp. 463–472. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-42111-7_36
3. Stankova, E.N., Ismailova, E.T., Grechko, I.A.: Algorithm for processing the results of cloud convection simulation using the methods of machine learning. In: Gervasi, O., et al. (eds.) ICCSA 2018. LNCS, vol. 10963, pp. 149–159. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-95171-3_13
4. Stankova, E.N., et al.: OLAP technology and machine learning as the tools for validation of the numerical models of convective clouds. Int. J. Bus. Intell. Data Min. **14**(1–2), 254–266 (2019)
5. Düntsch, I., Gediga, G.: Rough set data analysis. In: Encyclopedia of Computer Science and Technology (2000)
6. Pawlak, Z.: Why rough sets? In: Proceedings of IEEE 5th International Fuzzy Systems, vol. 2. IEEE (1996)
7. Shichkina, Y.A., Degtyarev, A.B., Koblov, A.A.: Technology of cleaning and transforming data using the knowledge discovery in databases (KDD) technology for fast application of data mining methods. CEUR Workshop Proceedings, vol. 1787, pp. 428–434 (2016)
8. Chapman, A.D.: Principles of data quality. GBIF (2005)
9. Fernández-Delgado, M., et al.: Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. **15**(1), 3133–3181 (2014)
10. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
11. Luo, G.: A review of automatic selection methods for machine learning algorithms and hyper-parameter values. Netw. Model. Anal. Health Inform. Bioinform. **5**(1), 18 (2016)
12. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Trans. Neural Networks **16**(3), 645–678 (2005)
13. Bagnall, A., et al.: The great time series classification bake off: an experimental evaluation of recently proposed algorithms (2016). Extended version. arXiv preprint arXiv:1602.01711
14. Fawaz, H.I., et al.: Deep learning for time series classification: a review (2018). arXiv preprint arXiv:1809.04356
15. Studer, R., Benjamins, V.R., Fensel, D.: Knowledge engineering: principles and methods. Data Knowl. Eng. **25**(1), 161–198 (1998)
16. Bernstein, A., Provost, F., Hill, S.: Toward intelligent assistance for a data mining process: an ontology-based approach for cost-sensitive classification. IEEE Trans. Knowl. Data Eng. **17**(4), 503–518 (2005)
17. 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
18. Tianxing, M., Zhukova, N.: An ontology of machine learning algorithms for human activity data processing. Learning **10**, 12 (2018)