ISSN: 1803-3814

NUMERICAL DATA CLUSTERING ONTOLOGY APPROACH

Peter Grabusts

Rezekne Academy of Technologies
Faculty of Engineering
Atbrivoshanas alley 115, Rezekne, LV-4601
Latvia
peteris.grabusts@rta.lv

Abstract: *Clustering algorithm tasks are used to group given objects defined by a set of numerical properties in such a way that the objects within a group are more similar than the objects in different groups. All clustering algorithms have common parameters the choice of which characterizes the effectiveness of clustering. The most important parameters characterizing clustering are: metrics, number of clusters and cluster validity criteria. In classic clustering algorithms semantic knowledge is ignored. This creates difficulties in interpreting the results of clustering. At present, the use of ontology opportunities is developing very rapidly, that provide an explicit model for structuring concepts, together with their interrelationship, which allows you to gain knowledge of a particular data model. According to the previously obtained results of clustering study, the author will make an attempt to create ontology-based concept from numerical data using similarity measures, cluster numbers, cluster validity and others characteristic features. To scientific novelty should be attributed the combination of approaches of classical data analysis and ontological approach to their structuring, that increases the efficiency of their use in engineering practice.*

Keywords: *Clustering, Cluster analysis, Ontology*.

1 Introduction

Nowadays there is a large amount of data in various fields of science, business, economics and other spheres and there is a need to analyze them for better management of a particular industry. Often, business needs stimulate to develop new, intelligent, data analysis methods that are oriented to practical application. The goal of cluster analysis as one of the basic tasks of intellectual data analysis is to search for independent groups (clusters) and their characteristics in analytical data [1], [2], [3]. Solving this problem allows to understand the data better, since clustering can be used in practically any application area where data analysis is required.

Author's research interests are oriented to clustering analysis: clustering algorithms, fuzzy clustering, rule extraction from clustered data etc. [4], [5]. The next step in the research is the implementation of ontologies in cluster analysis [6].

In order to evaluate the efficiency aspects of clustering activity, the following aim was put forward - to analyze and summarize the possibilities of clustering algorithms with the purpose of creating an ontology concept for numerical data clusterization.

Research tasks are subordinated to the aim stated:

- evaluate the validity of the choice of the metric;
- characterize the changes in the number of clusters to the analyzed data;
- evaluate the reliability of clustering results (clustering validity);
- get rules from clusters.

According to the previously obtained results of clustering study, the author will make an attempt to create ontology-based prototype of clustering concepts using similarity measures, cluster numbers, cluster validity and others characteristic features.

2 Clustering tasks and knowledge extraction

The cluster analysis is based on the hypothesis of compactness. It is assumed that the elements of the training set in the feature space are compact. The main task is to formally describe these formations. Clustering differs from the classification by the fact that there is no need to separate a variable group for analysis in the clustering process. From this point of view, clustering is treated as "non-teacher training" and is used at the initial stage of the research.

The cluster analysis is characterized by two features that distinguish it from other methods [2]:

- the result depends on the nature of the objects or their attributes, i.e. they can be uniquely identified objects or objects with a fuzzy description;
- the result depends on the possible relationship between the cluster and the objects in the clusters, i.e., the possible membership of the object in several clusters and the determination of the ownership of the object (strong or fuzzy membership) must be taken into account.

Taking into account the important role of clustering in the analysis of data, the concept of object ownership was generalized to the function of classes that determines the class objects belonging to a particular class. Two types of classes characterizing functions are distinguished:

- a discrete function that accepts one of the two possible values – belongs/does not belong to the class (classical clusterization);
- a function that accepts values from the interval [0,1]. The closer the values of the function to 1, the "more" the object belongs to a particular class (fuzzy clustering).

Clustering algorithms are mainly intended for the processing of multidimensional data samples, when the data is given in the table form "object-property". It allows to group objects in defined groups, in which objects are related to each other according to a particular rule. It does not matter how the following groups are called - taxon, clusters, classes, the main thing is that they accurately reflect the properties of these objects. After clustering, other methods of the intellectual data analysis are used for further analysis of data, in order to find out the nature of the acquired regularities and the possibilities of future use.

Clustering is commonly used in the data processing operation as a first step of analysis. It identifies similar data groups that can later be used to explore data interrelation. The cluster analysis process consists of the following steps:

- collection of necessary data for analysis;
- determination of characterizing sizes and boundaries for class data (clusters);
- data grouping into clusters;
- determination the hierarchy of classes and analysis of results.

All clustering algorithms have common parameters, the choice of which also characterizes the efficiency of the classification. The most important parameters characterizing clustering are: metrics (distance of cluster elements to the cluster center), number of clusters, evaluation of clustering reliability, possibility of obtaining the rules [7], [8], [9].

The author uses the [2] offered class clustering algorithm classification in his study.

Metrics. The main purpose of metrics learning in a specific problem is to learn an appropriate distance/similarity function. A metrics or distance function is a function which defines a distance between elements of a set [10], [11].

Euclidean distance is the most common use of distance – it computes the root of square differences between coordinates of a pair of objects.

Manhattan distance or city block distance represents distance between points in a city road grid. It computes the absolute differences between coordinates of a pair of objects.

Minkowski distance is the generalized metric distance.

Cosine distance is the angular difference between two vectors.

The summary of the metrics is shown in the Table 1.

Table 1: Distance measures and their applications

Measure	Examples and applications
Euclidean distance	K-means with its variations
Manhattan distance	Fuzzy ART, clustering algorithms
Cosine distance	Text Mining, document clustering

Traditionally Euclidean distance is used in the clustering algorithms, the choice of other metric in definite cases may be disputable. It depends on the task, the amount of data and on the complexity of the task.

Cluster numbers. An essential issue in implementing clustering algorithms is the determination of the number of clusters and initial centres. The simplest tasks assume that apriori is known the number of clusters and it is suggested to take the first m points of the training set as the initial values of the m cluster centres.

Clustering validity. Cluster validity is a method to find a set of clusters that best fits natural partitions (number of clusters) without any class information. There are three fundamental criteria to investigate the cluster validity: external criteria, internal criteria, and relative criteria [2].

External criteria. Given a data set X and a clustering structure C derived from the application of a certain clustering algorithm on X, external criteria compare the obtained clustering structure C to a pre-specified structure, which reflects a priori information on the clustering structure of X. For example, an external criterion can be used to examine the match between the cluster labels with the category labels based on apriori information.

Based on the external criteria, there is the following approach: comparing the resulting clustering structure C to an independent partition of the data P, which was built according to intuition about the clustering structure of the data set.

If P is a pre-specified partition of data set X with N data points and is independent of the clustering structure C resulting from a clustering algorithm, then the evaluation of C by external criteria is achieved by comparing C to P. Considering a pair of data points x_i and x_j of X, there are four different cases based on how x_i and x_j are placed in C and P.

Internal criteria. A Cophenetic correlation coefficient [12] is an index used to validate hierarchical clustering structures. Given the proximity matrix $P = \{p_{ij}\}$ of X, Cophenetic correlation coefficient measures the degree of similarity between P and the cophenetic matrix $Q = \{q_{ij}\}$, whose elements record the proximity level where pairs of data

points are grouped in the same cluster for the first time. The value of Cophenetic correlation coefficient lies in the range of [− 1, 1], and an index value close to 1 indicates a significant similarity between *P* and *Q* and a good fit of hierarchy to the data.

Relative criteria. External and internal criteria require statistical testing, which could become computationally intensive. Relative criteria eliminate such requirements and concentrate on the comparison of clustering results generated by different clustering algorithms or the same algorithm but with different input parameters.

The author uses External criteria in his research.

Rule extraction. The possibility of direct transforming clustering information into a symbolic knowledge frame is through rule extraction. Such assumptions are defined as *IF ... THEN ...* rules [9]. The benefits of rule extraction are as follows:

- the opportunity to test the acquired rules on different input data options is provided;
- limitations in the training data can be identified, thus clustering can be improved by introducing or removing additional clusters;
- determination of previously unknown regularities in data that is currently of growing importance in the Data Mining industry;
- a rule base can be created from the acquired rules, which could be used in future for similar types of applications.

Several neural networks use clustering in the training process, which results in the creation of hidden units, which are in fact cluster centres [9], [13]. The nature of each hidden unit enables a simple translation into a single rule:

$$IF\ Feature_1\ is\ TRUE\ AND\ IF\ Feature_2\ is\ TRUE\ ...\ AND\ IF\ Feature_n\ is\ TRUE$$
$$THEN\ Class_x$$

where a *Feature* is composed of upper and lower bounds calculated by the center μ_n positions, width σ and feature steepness *S*. The values of μ and σ are determined by the training algorithm. The upper and lower bounds are calculated as follows:

$$X_{lower} = \mu_i - \sigma_i + S \text{ and } X_{upper} = \mu_i + \sigma_i - S \quad (1)$$

Then rule extraction RULEX process can be seen below in Table 2 [9].

Table 2: Rule extraction algorithm

Procedure
For each hidden unit:
For each μ_i
$X_{lower} = \mu_i - \sigma_i + S$
$X_{upper} = \mu_i + \sigma_i - S$
Build rule by:
$antecedent = [X_{lower}, X_{upper}]$
Join antecedents with *AND*
Add class label
Write rule

Consequently, a base for the rules has been obtained.

3 Possibilities of ontologies

In recent years, the development of ontologies - an explicit formal description of the terms of the subject area and the relations between them - passes from the world of laboratory work through the artificial intelligence to the working desks of experts in subject areas. Ontologies have become commonplace in the World Wide Web. Ontologies in the network range from large taxonomies, categorizing websites to categorizations of merchandise traded and their characteristics. In many disciplines, standard ontologies are now being developed, which can be used by domain experts for general use and to annotate information in their field.

Informally, an ontology is a kind of description of the world in relation to a particular area of interests. This description consists of terms and rules for the use of these terms limiting their values within a particular area. On the formal level, an ontology is a system consisting of a set of concepts and a set of statements about these concepts on the basis of which it is possible to build classes, objects, relations, functions and theories.

The main components of the ontology are:

- Classes or concepts;
- Relations;
- Functions;

- Axioms;
- Examples.

There are various definitions of ontology but following definition has been generally accepted: "An ontology is a formal explicit specification of a shared conceptualization" [14]. Ontologies are often equated with taxonomic hierarchies of classes.

Thus, the goal of ontology is to accumulate knowledge in a general and formal way.

Ontologies can be classified in different forms. One of the most popular types of classification was offered by Guarino who classified types of ontologies according to their level of dependence on a particular task or point of view [15]:

- Top-level ontologies: describe general concepts like space, time, event, which are independent of a particular problem or domain.
- Domain-ontologies: describe the vocabulary related to a generic domain by specializing the concepts introduced in the top-level ontology.
- Task ontologies: describe the vocabulary related to a generic task or activity by specializing the top-level ontologies.
- Application ontologies: they are the most specific ones. Concepts often correspond to roles played by domain entities. They have a limited reusability as they depend on the particular scope and requirements of a specific application.

Ontologies are widely used in Semantic Web and document clustering, but there is very little information available about the use of ontologies in the numerical data clustering.

Thus, an ontology is an explicit representation of knowledge. It is a formal, explicit specification of shared conceptualizations, representing the concepts and their relations that are relevant for a given domain of discourse [14].

4 The concept of the ontology of numerical data clustering

The concept of the ontology of numerical data clustering, which is newly developed, consists of the following classes:

Class *Clustering_Task*. It is an abstract class. It is related to the clustering algorithm class. Depending on the purpose of the clustering and the domain of operation, the clustering algorithm, number of clusters and sample data are selected.

Class *Clustering_Algorithm*. This class represents a list of available clustering algorithms and their features (see Fig.1).

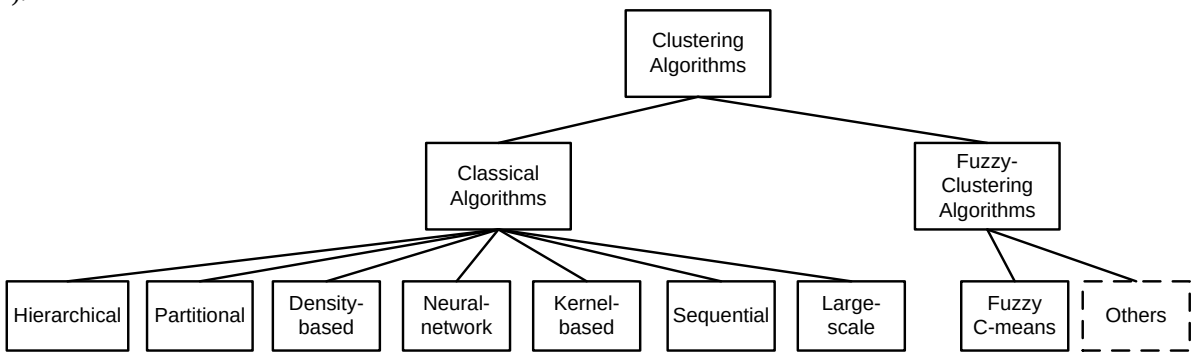

Figure 1: A hierarchical view of the *Clustering_algorithm* class

Class *Clustering_Metric*. This class represents a list of available distance metrics for clustering algorithms (see Fig.2).

Class *Clustering_Validity*. This class represents a list of cluster validity methods (see Fig.3).

Class *Clustering_Rule*. This class represents a list of rule extraction methods from clusters (if it is possible).

Clustering ontology concept should work according to the following scheme: numerical data selection, choice of clustering algorithm, determining the number of clusters, performance of clustering, validation of clustering, acquisition of rules (if possible) (see. Fig. 4.).

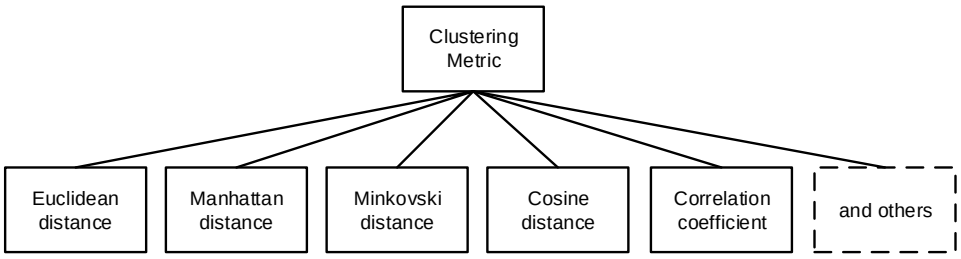

Figure 2: A hierarchical view of the *Clustering_Metric* class

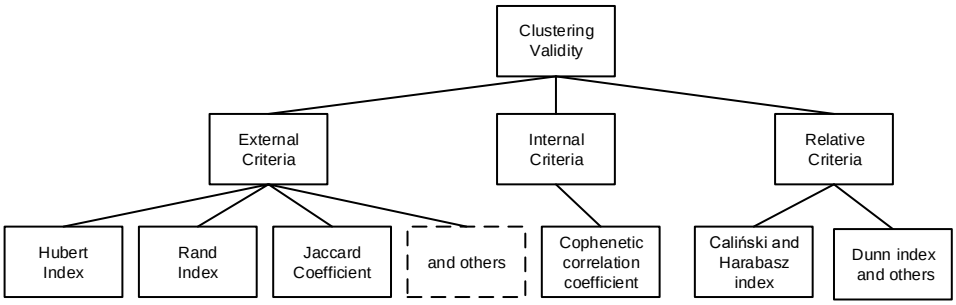

Figure 3: A hierarchical view of the *Clustering_Validity* class

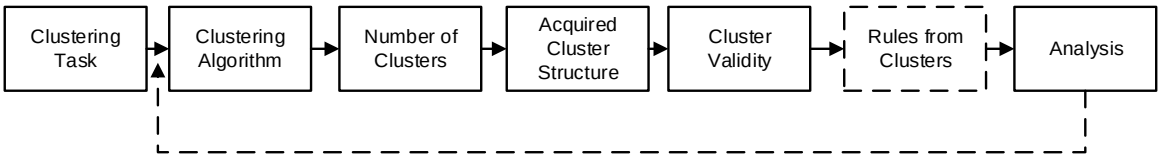

Figure 4: Clustering concept

Developing framework Protégé OWL tool is used for construct this concept [16].

Protégé is a special tool that is suitable for creating and editing ontologies, but OWL (Web Ontology Language) is the language by which it is possible to define ontologies. OWL ontology may include class descriptions, their properties and instances. The formal semantics of OWL describes how using this data obtain information that was not openly described in ontology but was derived from the data semantics.

Protégé is a free open-source platform that, with the help of a special tool set, allows to construct domain-based models and knowledge-based applications based on ontologies. Several knowledge modeling structures and activities supporting the creation, visualization and editing of ontologies in various image formats have been implemented in the Protégé environment.

The development of ontologies with Protégé begins with the definition and description of the class hierarchy, then instances of these classes and different types of relationships (properties in Protégé) are set in order to place more meaningful information within the ontology.

To demonstrate the development of ontology, four classes are taken: *Clustering-Algorithm*, *Clustering-Metric*, *Clustering-Validity* and *Clustering-Numbers* (see Figures 5-8).

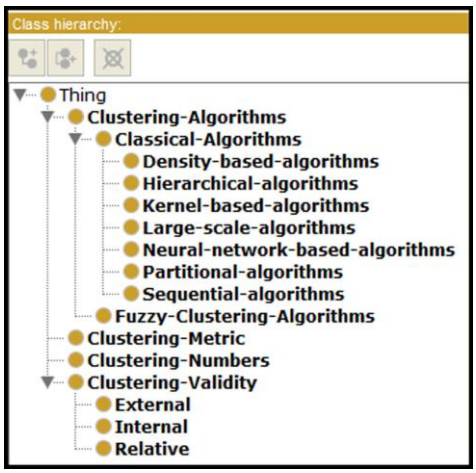

Figure 5: Clustering domain subclasses in the "Class hierarchy" tab of Protégé

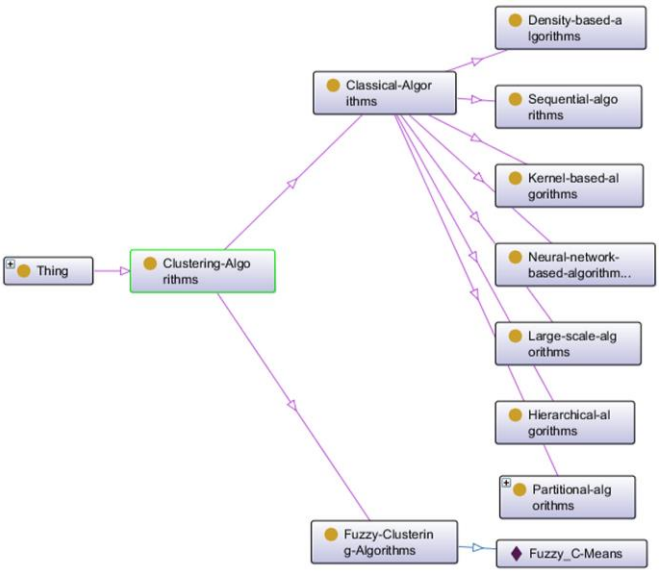

Figure 6: A visualization of the Class "Clustering subclasses" in OntoGraf tab

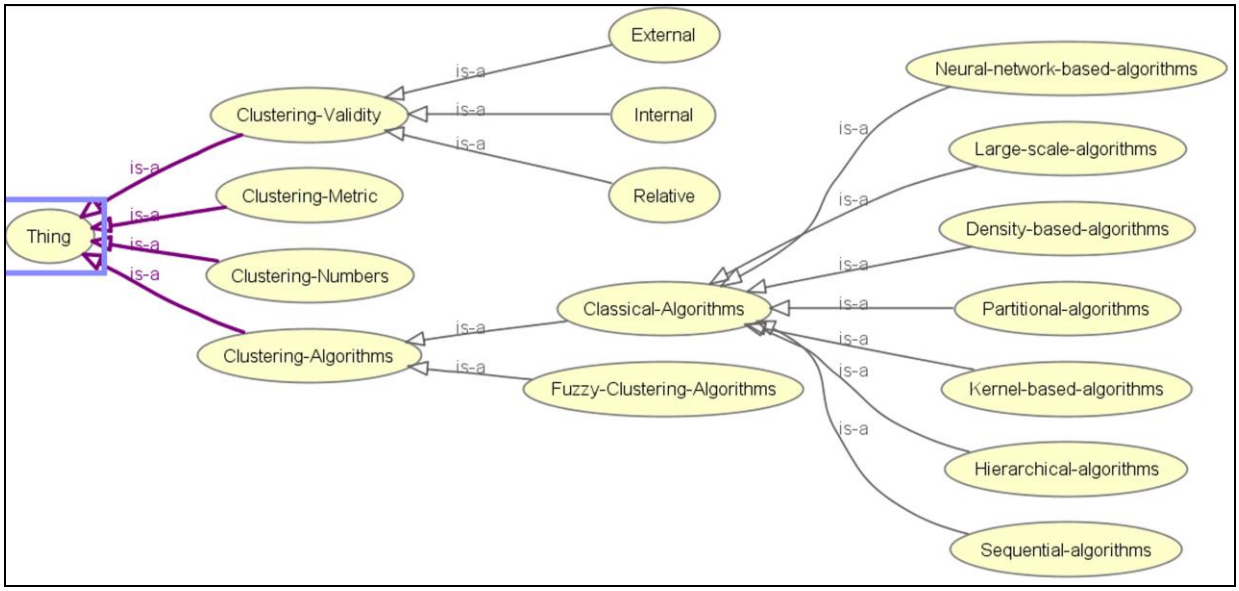

Figure 7: Clustering domain subclasses in the "OWLViz" tab of Protégé

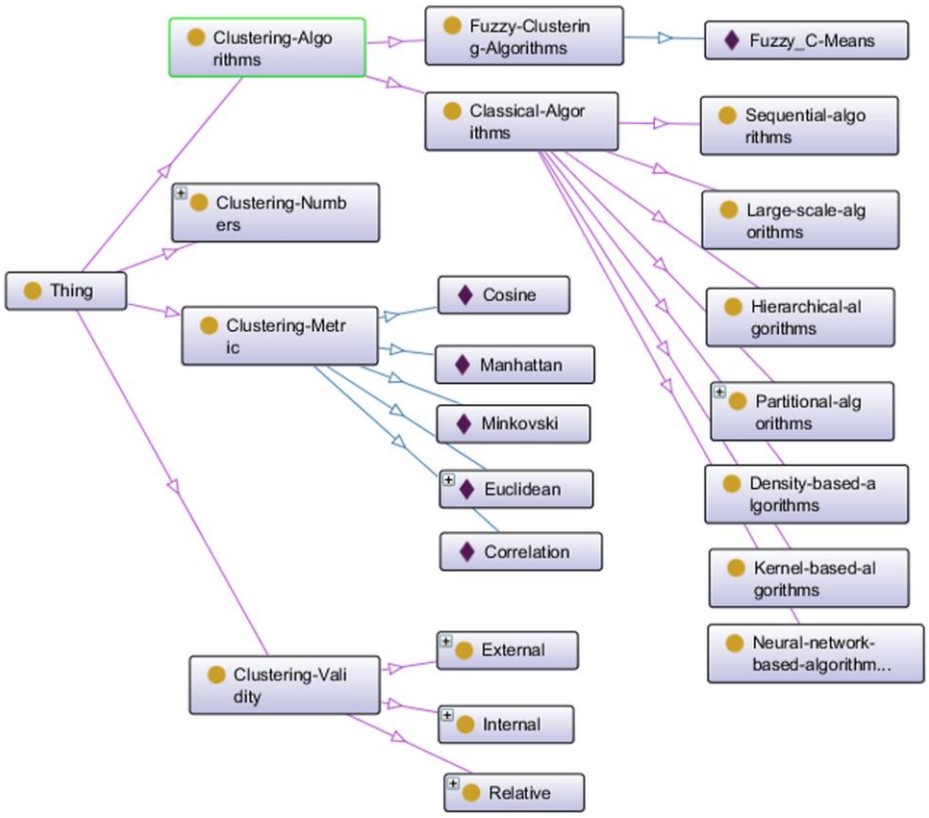

Figure 8: A visualization of the "Clustering" domain subclasses in OntoGraf tab

Since the clustering algorithm refers to the partition algorithm class, *K_Means* member was included in the *Partitional_algorithms* class. The K-means algorithm can use the metrics Euclidean distance or Manhattan distance; then the *Clustering_Metric* class include members: *Euclidean_distance* and *Manhattan_distance*.

The following properties were defined for member *K_Means*:

K_Means – use -> Euclidean_distance
K_Means – use -> Manhattan_distance.

In turn, the *Clustering_Metric* object *Euclidean_distance* is assigned a property
Euclidean_distance – isUsedBy -> K_means (see Fig. 9).

The Protege developer has recently offered a new product - WebProtégé, where the users can process their OWL data. There is currently no way to visualize the ontology as it was done with OntoGraf. It's now possible to work with Classes, Properties, Individuals.

An example of a demonstration clearly shows that Protégé can create an effective ontology description, but it is a sufficiently time consuming process. The author plans to continue the work on the further development of numerical data clustering ontology.

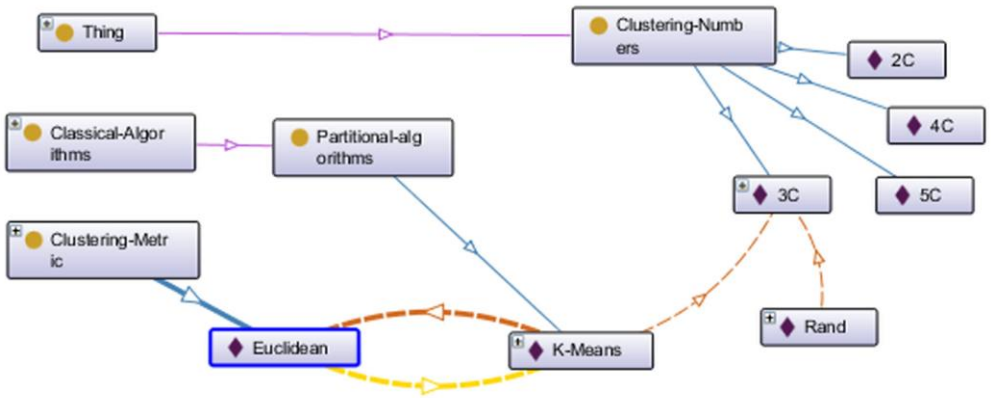

Figure 9: K-means property visualization in *Clustering_Metric* class

5 Conclusion

There are no directly formalized criterions in cluster analysis, therefore different clustering parameters are chosen in a subjective evaluation. This refers to the choice of the clustering algorithm, the choice of the number of clusters in each particular case, the determination of cluster validation criteria. Equally important is the acquisition of knowledge from clusters in the form of rules. All this creates problems in the interpretation of clustering results. In recent decades, cluster analysis has transformed from a single data analysis section in a separate direction that is closely related to knowledge support systems. Partly it has happened because of the introduction of concepts of ontology in the description of clustering characteristics. The use of clustering ontologies in document and semantic web applications is developing very rapidly, but the clustering of numerical data is neglected. The author has tried to formulate and create an ontology-based prototype for numerical data clustering. This concept contains several concept classes: clustering algorithms, cluster numbers, cluster validity, and other characteristics features. Further studies will focus on the clarification of these classes and developing a real-world model according to data clustering purposes. To scientific novelty should be attributed the combination of approaches of classical data analysis and ontological approach to their structuring, that increases the efficiency of their use in engineering practice.

References

[1] Everitt, B.S.: Cluster analysis. John Viley and Sons, London (1993)
[2] Xu, R., Wunsch, D.C.: Clustering. John Wiley & Sons (2010)
[3] Rui, X., Wunsch, D.: Survey of clustering algorithms. Neural Networks, IEEE Transactions, 16(3), 645–678 (2005)
[4] Hoppner, F., Klawonn, F., Kruse, R., Runkler, T.: Fuzzy Cluster Analysis. John Whiley and Sons, New York (1999)
[5] Crawen, M., Shavlik, J.: Using sampling and queries to extract rules from trained neural networks. In: Machine Learning: Proceedings of the Eleventh International Conference, San Francisco, CA (1994)
[6] Gašević, D., Djurić, D., Devedžić, V.: Model driven architecture and ontology development. Springer-Verlag (2006)
[7] Gan, G., Ma, C., Wu, J.: Data clustering: Theory, algorithms and applications. ASA-SIAM series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA (2007)
[8] Kaufman, L., Rousseeuw, P.J.: Finding groups in data, *An introduction to cluster analysis*. John Wiley & Sons (2005)
[9] Andrews, R., Gewa, S.: RULEX and CEBP networks as the basis for a rule refinement system. In: J. Hallam et al, editor, Hybrid Problems, Hybrid Solutions. IOS Press (1995)
[10] Vitanyi, P.: Universal similarity. ITW2005, Rotorua, New Zealand (2005)
[11] Li, M., Chen, X., Ma, B., Vitanyi, P.: The similarity metric. IEEE Transactions on Information Theory, vol.50, No. 12, 3250-3264 (2004)
[12] Jain, A.K., Dubes, R.C.: Algorithms for clustering data. Prentice Hall PTR (1988)
[13] Hush, D.R., Horne, B.G.: Progress in Supervised Neural Networks. What's new since Lippmann?. IEEE Signal Processing Magazine, vol.10, No 1. 8-39 (1993)
[14] Gruber, T.R.: A translation approach to portable ontologies. Knowledge Acquisition, 5(2), 199-220 (1993)
[15] Guarino, N.: Formal Ontology in Information Systems. In: 1st International Conference on Formal Ontology in Information Systems, FOIS, Trento, Italy, IOS Press, 3-15 (1998)
[16] Protégé project homepage. https://protege.stanford.edu [Online, accessed 2018/03/31]