Knowledge-based ontology concept for numerical data clustering

Peter Grabusts
Rezekne Academy of Technologies
Rezekne, Latvia
peteris.grabusts@rta.lv

***Abstract*—Classical clustering algorithms are sufficiently well studied, they are used for grouping numerical data in similar structures - clusters. Similar objects are placed in the same cluster, different objects in another cluster. All of the classic clustering algorithms have common parameters, and successful selection of which also determines the clustering result. The most important parameters characterizing clustering are: clustering algorithm, metrics, initial number of clusters, criteria for clustering accuracy. In recent years, there has been a tendency towards the possibility of obtaining rules from clusters. Classical clustering algorithms do not apply semantic knowledge. It creates difficulties in interpreting the results of clustering. Presently, the use of ontology opportunities is developing very rapidly, that allows to gain knowledge about a certain data model. The paper analyzes the concept of ontology and prototype development for numerical data clusterization, which includes the most significant indicators characterizing clusterization. The aim of the work is to develop a concept for analyzing clustering data with the help of ontologies. As a result of the work, a study has been conducted on the use of ontologies in this type of tasks.**

***Keywords*—clustering, cluster analysis, ontology**

I. 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, the needs of business stimulate to develop new intelligent methods for data analysis that are oriented towards practical application. The goal of cluster analysis as one of the basic tasks of intellectual data analysis - to search for independent groups (clusters) and their characteristics in analytic data [1], [2], [3]. Solving this problem allows to understand the data in a better way since clustering can be used in any application area where data analysis is required.

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

In order to evaluate the efficiency aspects of clustering, the aim was set - to analyze and summarize the possibilities of clustering algorithms with the purpose of creating an ontology prototype for numerical data clusterization. Research tasks are subjected to the stated aim:

- To review clustering algorithms.
- Carry out the evaluation of the eligibility of the metrics selection.
- Characterize the impact of changes in the number of clusters.
- Evaluate the reliability of the results of clustering (clusters validity).
- Evaluate the possibility to get rules from clusters.
- Develop the ontology concept for numerical data clustering.

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.

II. AN OUTLINE OF CLASSICAL CLUSTERING APPROACH

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 describe these formations formally. Clustering differs from the classification by the fact that there is no need to select a separate changeable 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 uniqualy certain objects or objects with a fuzzy description.
- The result depends on the nature of possible relationship between the cluster and the objects in the clusters, i.e., the possibility of belonging the object to several clusters and the determination of the ownership of the object (strong or fuzzy belonging) should be taken into account.

Taking into account the important role of clustering in the analysis of data, the concept of object belonging 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 function values are 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-attribute». They allow to group objects into certain groups, in which objects are related to each other according to a particular rule. It does not matter how the following groups are called - taxons, clusters or classes, the main thing is that they accurately reflect the properties of these objects. After the clustering, data for further analysis are used by other intellectual data analysis methods, 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 process as the first step of analysis. It identifies similar data groups that can later be used to explore data interrelation. The cluster analysis process formally consists of the following steps (see Fig. 1):

- Selection of necessary data for analysis.
- Determination of characterizing class sizes and boundaries for class data (clusters).
- Data grouping in clusters.
- Determining the class hierarchy and analyzing the results.

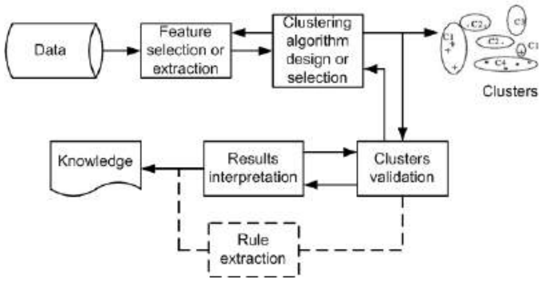


Figure 1. Clustering procedure

All clustering algorithms have common characteristics, the selection of which is characterized by a clustering efficiency. The most important clustering parameters are as follows: metric (cluster element distance to the cluster center), the number of clusters k, clustering validity assessment, opportunity to get rules [7], [8], [9].

III. Clustering characteristics

Classes of clustering algorithms are shown in the Fig. 2:

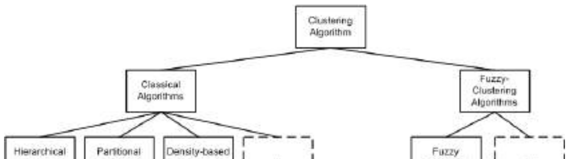


Figure 2. Hierarchical view of the clustering algorithm class

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]. A set with a metric is called a metric space. In many data retrieval and data mining applications, such as clustering, measuring similarity between objects has become an important part. In general, the task is to define a function Sim(X,Y), where X and Y are two objects or sets of a certain class, and the value of the function represents the degree of "similarity" between the two.

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 distance measure can also be derived from the correlation coefficient, such as the Pearson correlation coefficient. Correlation coefficient is standardized angular separation by centering the coordinates to its mean value. It measures similarity rather than distance or dissimilarity.

The summary of the metrics is shown in the Fig. 3.

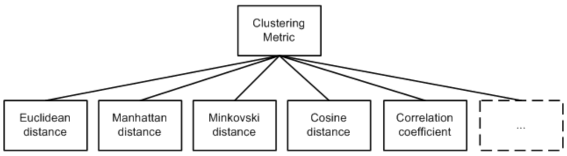


Figure 3. Hierarchical view of the clustering metrics class

Traditionally Euclidean distance is used in 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 of clustering algorithms is the determination of the number of clusters and initial centers. In the simplest tasks it is assumed that the number of clusters is known apriori and it is suggested to take the first m points of the training set as the initial values of the m cluster centers.

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. In this case only external cluster validity index was analyzed [2].

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 prespecified 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.

The most popular clustering quality external criteria are shown in the Fig. 4.

For example, Rand index suggests an objective criterion for comparing two arbitrary clusterings based on how pairs of data

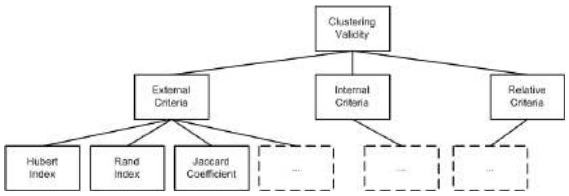


Figure 4. Hierarchical view of the clustering validity class

points are clustered. Given two clusterings, for any two data points there are two cases:

- The first case is that the two points are placed together in a cluster in each of two clusterings or they are assigned to different clusters in both clusterings.
- The second case is that the two points are placed together in a cluster in one clustering and they are assigned to different clusters in the other.

Rule extract. The possibility of direct transforming clustering information into a symbolic knowledge form is through rule extraction. Such assumptions are given as IF ... THEN ... rules [12]. The benefits of rule-making are as follows:

- The opportunity to check the acquired rules for different input data options is provided.
- Deficiencies in the training data can be identified, thus clustering can be improved by introducing or removing additional clusters.
- Determination of previously unknown regularities in the Data Mining industry.
- The rule base can be created from the acquired rules, which could be used in future for similar types of applications.

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

IF Feature1 is TRUE AND IF Feature2 is TRUE ... AND IF FeatureN is TRUE THEN ClassX.

Thus, the knowledge base is collected for the most important indicators characterizing clustering.

IV. Ontology based approach

In this paper author presents a formal clustering ontology framework concept, which can provide the background for numerical data clustering. Using the ontology, numerical clustering can become a knowledge-driven process.

As it was mentioned in the previous chapters clustering is using at the data level instead of the knowledge level, which helps from precisely identifying targets and understanding the clustering results. Existing clustering methods consider various constraints and they only consider limited knowledge concerning the domain and the users. In such a way, to include domain knowledge in the clustering methods and clustering process becomes an important topic in clustering data research and analysis.

There are many and different definitions of ontologies, but the following definition has recently been accepted as generally recognized: „An ontology is a formal explicit specification of a shared conceptualization" [14]. Ontologies are often equated with taxonomic hierarchies of classes. It can be said that the purpose 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 is offered by Guarino, who classified types of ontologies according to their level of dependence on a particular task or point of view [15].

It should be noted that ontologies are widely used in document clustering and Semantic Web, but undeservedly forgotten in 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].

The concept of the numerical data clustering ontology to be developed consists of the following classes:

Clustering-Task. It is an abstract class. This is connected to the clustering algorithm class. Depending on the purpose of the clustering and the domain, the clustering algorithm, the number of clusters and sample data are selected.

Clustering-Algorithm. This class represents a list of available clustering algorithms and their features (see Fig.2).

Clustering-Metric. This class represents a list of available distance metrics for clustering algorithms (see Fig.3).

Clustering-Numbers. This class represents a list of available numbers of clusters.

Clustering-Validity. This class represents a list of cluster validity methods (see Fig.4).

Clustering-Rule. This class represents a list of rule extraction methods from clusters (if it is possible).

Based on such class analysis the following approach is offered for ontology-based clustering, as shown in the Fig. 5.

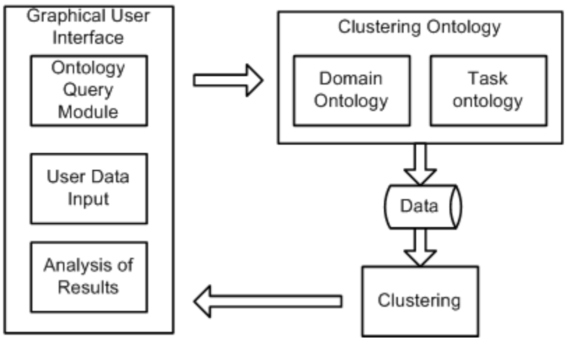


Figure 5. The framework concept of ontology-based numerical data clustering

Clustering ontology prototype 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. 6.).

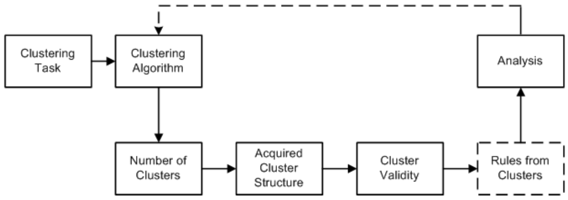


Figure 6. Working scheme of clustering prototype

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

Protege is an ontology and knowledge base editor. Protege is a tool that enables the construction of domain ontologies, customized data entry forms to enter data. Protege allows the definition of classes, class hierarchies, variables and the relationships between classes and the properties of these relationships.

Protege is a special tool, which is thought to create and edit ontology, but OWL (Web Ontology Language) is a language through which it is possible to define the ontology. OWL ontology may include descriptions of classes, their characteristics and their instances. OWL formal semantics describes how, using these data get information which was not openly described in ontology, but which follows from the data semantics. Protege is a free open-source platform, which contains special tool kit which makes it possible to construct domain models and knowledge-based applications based on ontologies. In Protege environment a number of knowledge-modeling structures and actions that support ontology creation, visualization and editing of different display formats are implemented.

The development of ontologies with Protege begins with the definition and description of the classes hierarchy, after that the instances are assigned of these classes and different type of relationships (properties in Protege) in order to put more meaningful information within the ontology [17].

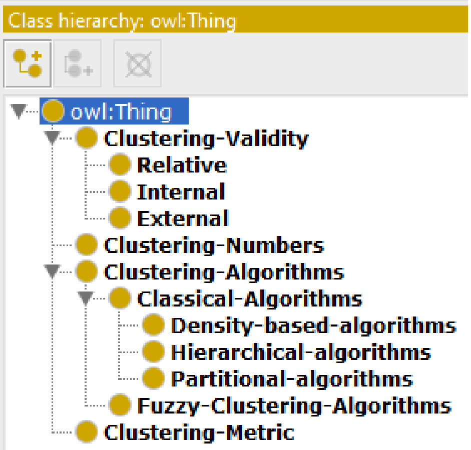
To demonstrate the development of ontology, four classes are taken: Clustering-Algorithm, Clustering-Metric, Clustering-Validity and Clustering-Numbers (see Fig. 7, Fig. 8 and Fig. 10).

Since the clustering algorithm refers to the partition algorithm class, K-Means member was included in the partition algorithms class. The K-means algorithm can use the metric Euclidean-distance or Manhattan-distance; then in Clustering-Metric class the members Euclidean-distance or Manhattan-distance are included.

The following properties were defined for K-Means:

K-Means – use -> Euclidean-distance or

K-Means – use -> Manhattan-distance


Class hierarchy: owl:Thing

owl:Thing
Clustering-Validity
Relative
Internal
External
Clustering-Numbers
Clustering-Algorithms
Classical-Algorithms
Density-based-algorithms
Hierarchical-algorithms
Partitional-algorithms
Fuzzy-Clustering-Algorithms
Clustering-Metric

Figure 7. Clustering domain subclasses in the “Class hierachy” tab of Protege

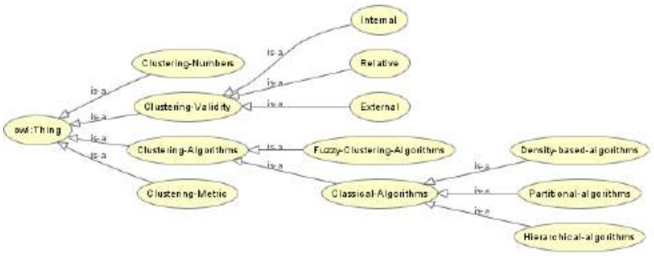


Figure 8. Clustering domain subclasses in the “OWLViz” tab of Protege

To simplify, we assume that three clusters are being studied in clustering and validation testing is carried out using the Randa index. Accordingly, we obtain:

K-Means – use -> 3C

Rand – use -> 3C

In turn, the Clustering-Metric object Euclidean-distance is assigned a property

Euclidean-distance – isUsedBy -> K-means

3C – isUsedBy -> K-means

3C – isUsedBy -> Rand

The result is shown in the Fig. 11.

Recently, the Protege developer has offered a new product-WebProtege, where the users can process their OWL data. WebProtege is an ontology development environment for the Web that makes it easy to create, upload, modify, and share ontologies for collaborative viewing and editing. WebProtege is lightweight ontology editor and knowledge acquisition tool for the Web [16].

Currently, there is no possibility to visualize the ontology as it was done with OntoGraf. Now it is possible to work with Classes, Properties, Individuals (see Figure 9).

An example of a demonstration shows that with a help of Protege an effective ontology description can be created,

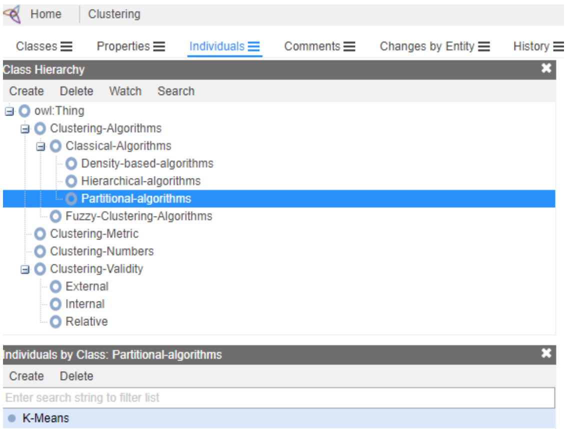


Figure 9. Clustering class hierarchy in WebProtege

but it is a sufficiently laborious process. The author plans to continue the work on the further development of numerical data clustering ontology.

V. Conclusion

There are no directly formalized criteria in the 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 causes a problem 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 linked to knowledge support systems. Partly it happened due to 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 undeservedly neglected. The author is striving 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. Future studies will focus on specification of these classes and developing a real model appropriately to the data clustering purposes.

References

[1] B. S. Everitt, *Cluster analysis*, John Viley and Sons, London, 1993.

[2] R.Xu and D.C.Wunch, *Clustering*, John Wiley and Sons, 2010, pp. 263-278.

[3] X. Rui and D. Wunsch II, Survey of clustering algorithms, *Neural Networks*, IEEE Transactions, 16(3):645–678, May 2005.

[4] F. Hoppner, F. Klawonn, R. Kruse and T. Runkler, *Fuzzy Cluster Analysis*, John Whiley and Sons, New York, 1999,289 p.

[5] M. Crawen and J. Shavlik, Using sampling and queries to extract rules from trained neural networks, *Machine Learning: Proceedings of the Eleventh International Conference*, San Francisco, CA, 1994.

[6] D. Gašević, D. Djurić and V. Devedžić, *Model driven architecture and ontology development*, Springer-Verlag, 2006.

[7] G. Gan, C. Ma and J. Wu, *Data clustering: Theory, algorithms and applications*, ASA-SIAM series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 2007.

[8] L. Kaufman and P. J. Rousseeuw, *Finding groups in data. An introduction to cluster analysis*, John Wiley and Sons, 2005.

[9] R. Andrews and S. Gewa, *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] P. Vitanyi, *Universal similarity*, ITW2005, Rotorua, New Zealand, 2005.

[11] M. Li, X. Chen, B. Ma and P. Vitanyi, The similarity metric, *IEEE Transactions on Information Theory*, vol.50, No. 12, pp.3250-3264, 2004.

[12] S. Russel and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 2010, 1132 p.

[13] D. R. Hush and B. G. Horne, Progress in Supervised Neural Networks. What's new since Lippmann?, *IEEE Signal Processing Magazine*, vol.10, No 1., p.8-39, 1993.

[14] T. R. Gruber, A translation approach to portable ontologies, *Knowledge Acquisition*, 5(2), 199-220, 1993.

[15] N. Guarino, Formal Ontology in Information Systems. *1st International Conference on Formal Ontology in Information Systems*, FOIS, Trento, Italy, IOS Press, 3-15, 1998.

[16] Protege project homepage. Available at: http://protege.stanford.edu/ (Accessed 2017 Nov).

[17] P. Grabusts, The Concept of Ontology for Numerical Data Clustering, *Proceedings of the 9th International Conference „Environment. Technology. Resources"*, Rezekne, Latvia, June,20-22, 2013, Vol.2., P.11-16.

КОНЦЕПЦИЯ ОНТОЛОГИИ ОСНОВАННАЯ НА ЗНАНИЯХ ДЛЯ КЛАСТЕРИЗАЦИИ ЧИСЛОВЫХ ДАННЫХ

Грабуст П.С.

Резекненская Академия Технологий, Латвия

Классические алгоритмы кластеризации достаточно хорошо изучены, они используются для группировки числовых данных в аналогичных структурах - кластерах. Аналогичные объекты помещаются в один кластер, разные объекты в другой кластер. Все классические алгоритмы кластеризации имеют общие параметры, и их успешный выбор также определяет результат кластеризации. Наиболее важными параметрами, характеризующими кластеризацию, являются: алгоритм кластеризации, метрика, начальное количество кластеров, критерии точности кластеризации. В последние годы наблюдается тенденция к возможности получения законов из кластеров. В классических алгоритмах кластеризации семантические знания не используются. Это создает трудности при интерпретации результатов кластеризации. В настоящее время использование возможностей онтологии очень быстро развивается, что позволяет получить знания об определенной модели данных. В данной работе проанализирована концепция онтологии и разработан прототип для кластеризации числовых данных, что включает в себя наиболее важные показатели, характеризующие кластеризацию. Целью работы является разработка концепции анализа данных кластеризации с помощью онтологий. В результате работы было проведено исследование о возможностях использования онтологий в задачах такого типа.

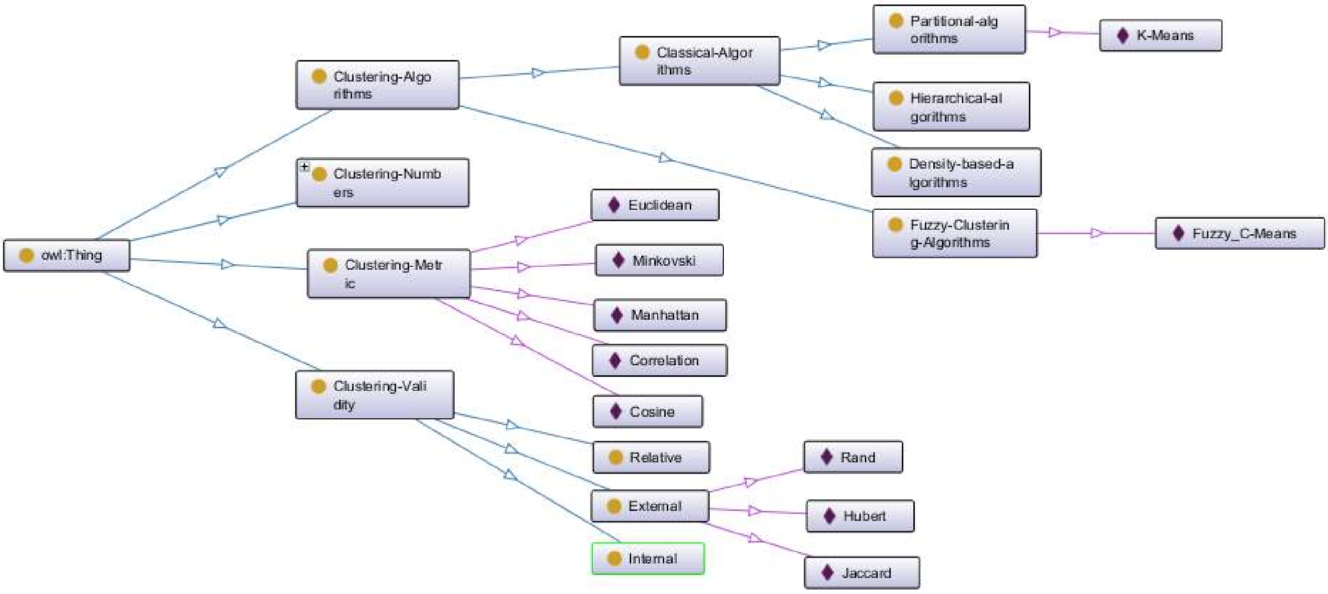


Figure 10. A visualization of the Clustering domain subclasses in OntoGraf tab

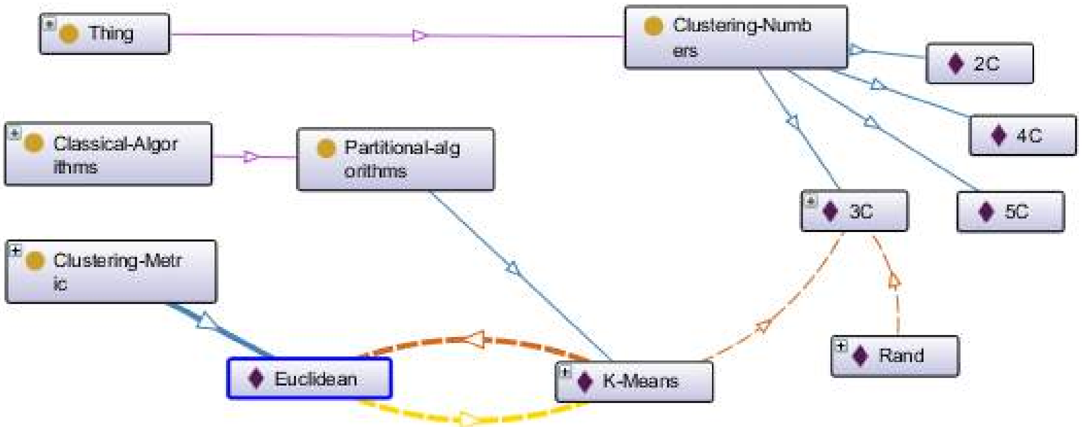


Figure 11. Visualization of K-means properties in the Clustering-Metric class