Analysis of ontology based approach for clustering tasks

Peter Grabusts

Engineering Faculty, Rezekne Higher Educational Institution, Atbrivoshanas alley 90, Rezekne, LV-4601, Latvia
peter@ru.lv

Abstract: *Clustering algorithms 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 k and cluster validity criteria. In classical clustering algorithms semantic knowledge is ignored. This creates difficulties in interpreting the results of clustering. Currently, the possibility to use ontology opportunities is developing rapidly, that provide an explicit model for structuring concepts, together with their interrelationship, that allows to gain knowledge on a specific data model. According to the previously obtained results of clustering study, the author will make a first attempt to create ontology based prototype of clustering concepts from numerical data using similarity measures, cluster numbers, cluster validity and others characteristic features.*

Keywords: clustering, cluster analysis, ontology.

Introduction

Nowadays there is a large amount of data accumulated in the various fields of science, business, economy, etc. areas and there is a need to analyze them for better management of one separate sector. Often business requirements stimulate the development of new intelligent data analysis methods that are focused on practical applications. Clustering is as one of the intelligent data analysis tasks and its aim is to search for an independent group (cluster) and its performance in the test data (Everitt et al., 1993). Resolving such a task leads to better understand of the data, because clustering can be used practically in any area of application 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 (Crawen et al., 1994; Hoppner et al., 1999). The next step in the research would be the implementation of ontologies in cluster analysis.

To evaluate the clustering performance aspects the following purpose was put forward – to analyze and summarize the clustering algorithms possibilities in order to create an ontological prototype for numerical data clustering. The work being carried out and following tasks were set:

- to carry out the evaluation of the validity of metrics chose;
- characterize the change in the number of clusters of analyzed data;
- evaluate the reliability of the results of clustering (clusters validity);
- extract the rules from the 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.

Clustering tasks and knowledge extraction

Clustering is based on the hypothesis of compactness. It is assumed that the set of elements of the learning characteristics of the space is in a compact way. The main task is a formalized description of these formations. Some clustering algorithms are well known, such as Isodata, FOREL, k-means etc.

Clustering differs from classification in following - in cluster analysis process there is no need to distribute a separate variable group. From this point of view, clustering is considered as a "learning without a teacher" and is used in the initial stage of the research.

Clustering is characterized by two features that distinguish it from other methods:

- the result depends on the object itself or the kind of attributes used, namely, they can be clearly defined objects or objects with fuzzy description;
- the result depends on the potential of the cluster and the object relations of natural clusters, that is, we should take into account the possible object belonging to multiple clusters and object ownership detection (strong or fuzzy membership).

Given the important role of clustering in data analysis, object ownership concept was generalized to a class function that defines the belonging of class object to proper class. Two classes of characteristic functions are distinguished:

- a discrete function that accepts one of two possible values - belong / not belong (classical clustering);

- a function that accepts values from the interval [0,1]. The closer the value of function is to 1, the more the subject belongs to a certain class (fuzzy clustering).

Clustering algorithms are mainly designed for multi-dimensional data sampling processing, when the data are given in tabular form in the "object-property". They allow you to group objects into groups, where objects relate to each other by a specific rule. It does not matter, how these groups are named- taxons, clusters, classes, as long as it reasonably reflects the properties of this object. After clustering the data for further analysis other intelligent data analysis techniques are used to determine the nature of the resulting regularities and future uses.
Clustering is typically used for data processing as a first step in the analysis. It identifies groups of similar data that can later be used for the investigation of relationships between the data. Clustering formal process consists of the following stages:

- collecting the required data for analysis;
- determination the characteristics of the clusters;
- data grouping in clusters;
- class hierarchy definition and analysis of the results.

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 (Gan et al., 2007; Kaufman et al., 2005).

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 (Li et al., 2004; Vitanyi, 2005). 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. Formally, a distance is a function D with nonnegative real values, defined on the Cartesian product X x X of a set X. It is called a metrics on X if for every x,y,z ϵ X:

- D(x,y)=0 if x=y (the identity axiom);
- D(x,y) + D(y,z) $\geq$ D(x,z) (the triangle inequality);
- D(x,y)=D(y,x) (the symmetry axiom).

A set X provided with a metric is called a metric space.
Euclidean distance is the most common use of distance – it computes the root of square differences between coordinates of a pair of objects:

$$D_{XY} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} \tag{1}$$

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:

$$D_{XY} = \sum_{k=1}^{d} | x_{ik} - x_{jk} | \tag{2}$$

Minkowski distance is the generalized metric distance:

$$D_{XY} = (\sum_{k=1}^{d} | x_{ik} - x_{jk} |^p)^{1/p} \tag{3}$$

Note that when p=2, the distance becomes the Euclidean distance. When p=1, it becomes city block distance.
Cosine distance is the angular difference between two vectors:

$$D_{XY} = \cos(\theta) = \frac{X \bullet Y}{\|X\| \|Y\|} = \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n} (X_i)^2} \times \sqrt{\sum_{i=1}^{n} (Y_i)^2}} \tag{4}$$

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 important issue in the implementation of clustering algorithm is the number of clusters and initial centers determination. In simple tasks it is assumed that *a priori* is known the number of clusters and as the initial cluster centers m values is offered to take the first set of training points m.

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 (Xu et al., 2009). In this case only external cluster validity index was analyzed.

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.

- Case 1: xi and xj belong to the same clusters of C and the same category of P.
- Case 2: xi and xj belong to the same clusters of C but different categories of P.
- Case 3: xi and xj belong to different clusters of C but the same category of P.
- Case 4: xi and xj belong to different clusters of C and different category of P.

Correspondingly, the numbers of pairs of points for the four cases are denoted as a, b, c and d. Because the total number of pairs of points is N(N-1)/2, denoted as M, we have:

$$M = a + b + c + d = \frac{n(n-1)}{2} \tag{5}$$

where n is the number of data points in the data set. When C and P are defined, one can choose one of the many clustering quality criteria. In the given research clustering quality criteria have been evaluated with the help of Rand or Hubert index (Xu et al., 2009).

Rand index is calculated by using the following formula:

$$R = \frac{a + d}{M} \tag{6}$$

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

Hubert index is calculated by using the following formula:

$$\Gamma = \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{ij} Y_{ij} \tag{7}$$

The value of both index ranges between 0 and 1. A higher index value indicates greater similarity between C and P.

Rule extraction. The possibility of directly converting clustering information in the form of symbolic knowledge extraction is through the rules (rule extraction). These assumptions are formulated as IF ... THEN ... rules (Russel et al., 2010). The benefits of the mining rules are as follows:

- the opportunity to verify the extracted rules on different variants of the input data is given;
- failures of training data can be identified, thus clustering operation can be improved by introducing new or removing additional clusters;
- determination of a previously unknown regularities in the data that currently have a growing importance of Data Mining industry;
- the resulting rules can be set up as a base of rules, which might also be used for similar types of applications.

Several artificial neural network algorithms use clustering during the learning process, leading to hidden units, which are, actually, cluster centers (Hush et al., 1993). The nature of each hidden unit enables a simple translation into a single rule:

$$\text{IF Feature}_1 \text{ is TRUE AND IF Feature}_2 \text{ is TRUE ... AND IF Feature}_n \text{ is TRUE}$$
$$\text{THEN Class}_x \quad (8)$$

where a *Feature* is composed of upper and lower bounds calculated by the center μ_n positions, width σ and feature steepness S. The value of the steepness was discovered empirically to be about 0.6 and is related to the value of the width parameter. The values of μ and σ are determined by the training algorithm. The upper and lower bounds are calculated as follows:

$$X_{lower} = \mu_i - \sigma_i + S \text{ and } X_{upper} = \mu_i + \sigma_i - S \quad (9)$$

Then rule extraction RULEX process can be seen below in Table 2 (Andrews et al., 1995).

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.

Ontology based approach

In this paper the 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 chapter, clustering is used at the data level instead of the knowledge level, that helps with identifying targets precisely 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 different definitions of ontology but the most common is recognized the following: an ontology is a formal explicit specification of a shared conceptualization (Gruber, 1993). Ontologies are often equated with taxonomic hierarchies of classes. It can be said, that the aim of ontology is to accumulate knowledge in general and formal way.

Ontologies can be classified into 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 (Guarino, 1998):

- *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.

It should be noted that ontologies are widely used in document clustering and Semantic Web but numerical data clustering is undeservedly forgotten.
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 (Gruber, 1993).
The newly developed numerical data clustering otology concept is composed of the following classes:

Clustering_Task. It is an abstract class. It is related to the proper clustering algorithm class. Depending on the purpose and the clustering area (domain), the clustering algorithm, the number of clusters and the data samples are chosen.

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

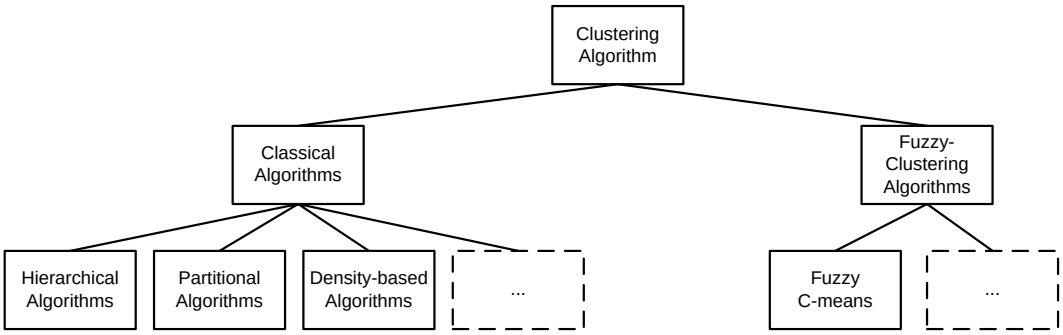

Fig. 1. **A hierarchical view of the *Clustering_Algorithm* class.**

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

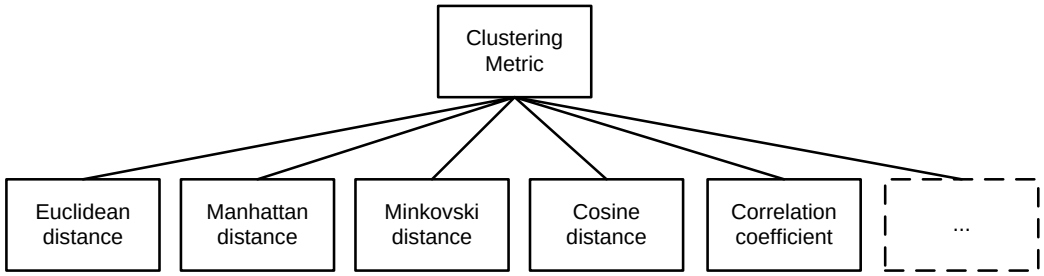

Fig. 2. **A hierarchical view of the *Clustering_Metric* class.**

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

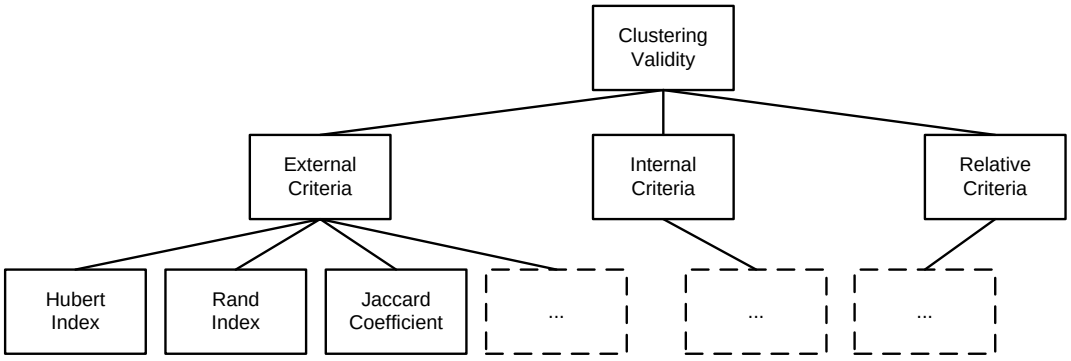

Fig. 3. **A hierarchical view of the *Clustering_Validity* class.**

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 Fig. 4.

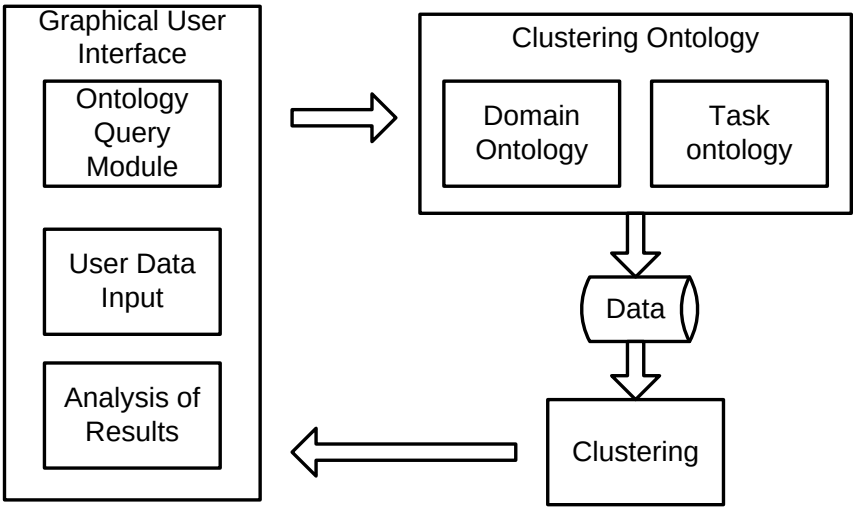

Fig. 4. **The framework concept of ontology-based numerical data clustering.**

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

Experimental part: bancruptcy data analysis

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) (Fig. 5).

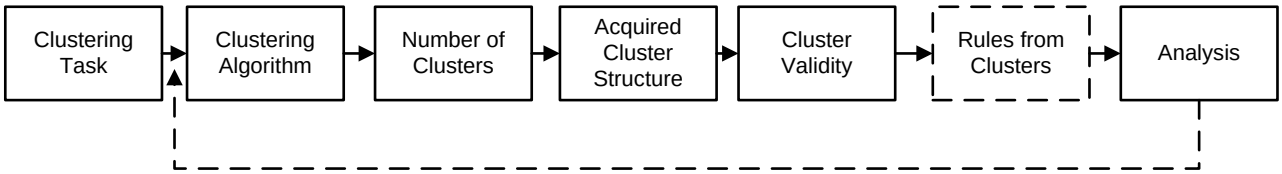

Fig. 5. **Working scheme of clustering prototype.**

The data on firm bankruptcy were taken from (Rudorfer, 1995). For the purpose of experiments, balance sheet data of 63 companies were used (46 - bankruptcy and 17 - not bankruptcy). It was decided to calculate the following financial ratios on the basis of the data available and further use them in all the experiments (Altman, 1968):

- R3: Cash Flow / Total Assets;
- R7: Current Assets / Current Liabilities;
- R9: Current Assets / Total Assets;
- R31: Working capital / Total assets.

The use of clustering algorithm k-means for this data set showed the following results: 59 – bankruptcy and 4 – not bankruptcy.

In order to verify clustering validity, quality index has been calculated – Rand and Hubert index for five clusters. Cluster structure C (consecutively with the number of clusters between 2 and 5 clusters) has been compared with specified divisions P containing various possible clusters.

Further, the total error has been calculated. The following errors of overall clustering have been calculated: for 2 clusters – 66.67 % (error rate of cluster 1: 71.19 %, error rate of cluster 2: 0.00 %).

Among all structures the lowest mistake occurs with 2 clusters, namely, 2 cluster structure in this case is the most optimal. Fig. 6 shows the calculated Rand and Hubert index for 2 cluster structure.

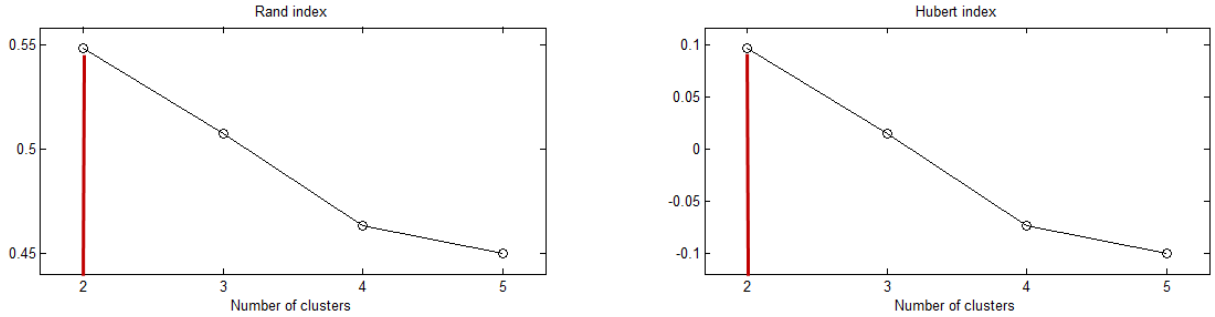

Fig. 6. **Rand and Hubert index**.

It should be noted that for this experiment, the data sample selected is not very successful, but nevertheless it illustrates the principles of clustering algorithm work. Indexes characterizing the quality of clustering are useful for analyzing the performance of clustering algorithms. With their help it is possible to choose an optimal cluster structure in cases when data distribution in clusters has not initially been set.

The objective of the next experiment was to extract rules from the bankruptcy data. Table 3 shows the results obtained in the course of the experiment whereas Table 4 lists the rules obtained under separate S values (Cluster 1 contains data on not bankrupt but Cluster 2 – data on bankrupt companies).

From Table 3 it can be seen that the rules obtained correctly describe bankruptcy data (44 out of 46) within the whole domain of parameter S, i.e., it can be stated that bankruptcy data are located in a fairly compact class.

Table 3

Results of bankruptcy data set training

	Values of parameter S								
Correct	-0.9	…	0	0.1	0.2	0.3	0.4	0.5	0.6
Cluster 1	15		12	10	9	6	5	1	0
Cluster 2	44		44	44	44	44	44	44	44
%	93.7	…	88.9	85.7	84.1	79.4	77.8	71.4	69.8

Table 4

Bankruptcy data set: characteristics of the extracted rules

	Parameter S= -0.9	**Parameter S= 0.4**
Values of centres and radii	Class 1= 0.03 1.25 0.74 0.01 Class 2= 0.13 1.86 0.59 0.10 Values of radii = 0.68 3.92	Class 1= 0.03 1.25 0.74 0.01 Class 2= 0.13 1.86 0.59 0.10 Values of radii = 0.68 3.92
Rules correctly describe elements of classes (%)	93.7	77.8
Rule of Cluster 1	IF (X1>= -1.54 AND < 1.61) AND IF (X2>= -0.33 AND < 2.83) AND IF (X3>= -0.84 AND < 2.32) AND IF (X4>= -1.57 AND < 1.59) THEN NON-BANKRUPT	IF (X1>= -0.24 AND < 0.31) AND IF (X2>= 0.97 AND < 1.53) AND IF (X3>= 0.46 AND < 1.02) AND IF (X4>= -0.27 AND < 0.29) THEN NON-BANKRUPT
Rule of Cluster 2	IF (X1>= -4.69 AND < 4.95) AND IF (X2>= -2.97 AND < 6.68) AND IF (X3>= -4.23 AND < 5.41) AND IF (X4>= -4.72 AND < 4.93) THEN BANKRUPT	IF (X1>= -3.39 AND < 3.65) AND IF (X2>= -1.67 AND < 5.38) AND IF (X3>= -2.93 AND < 4.11) AND IF (X4>= -3.42 AND < 3.63) THEN BANKRUPT

Conclusion

In clustering there is no directly formalized criterion, so different clustering parameters are chosen by subjective assessment. This refers to the selection of clustering algorithm and the number of clusters in each case, also to the cluster validation criteria. Also it is very important to get knowledge from clusters in the form of rules. All this leads to some problems in interpreting the results of clustering. In recent decades cluster analysis has been transformed from one of the data analysis sections in a separate direction, which is closely related to knowledge support system. Partly, this has happened due to the introduction of the ontology concept in the clustering characteristics description. The use of clustering ontology for documents and Semantic Web applications is

expanding rapidly but the numerical data clustering is undeservedly neglected. The author has made an attempt to define and develop an ontology-based prototype for clustering numerical data. This concept contains several concept classes: clustering algorithms, numbers of clusters, cluster validity and other characteristic features. In future studies these classes refinement will be carried out, also the real model according to data clustering purpose will be worked up.

References

Altman, E., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. In:*Journal of Finance*, vol. 13, pp.589-609.

Andrews, R. and Gewa, S., 1995. RULEX and CEBP networks as the basis for a rule refinement system. In: *J. Hallam et al, editor, Hybrid Problems, Hybrid Solutions*. IOS Press.

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

Everitt, B.S., 1993. Cluster analysis. *John Viley and Sons*, London, 170 p.

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

Gruber, T. R., 1998. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2) (1993) 199-220.

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

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

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

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

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

Protégé project homepage: http://protege.stanford.edu/index.html , 05.03.2013.

Rudorfer, G., 1995. Early bankruptcy detecting using neural networks. In:*APL Quote Quad*, ACM New York, vol. 25, N. 4, P. 171-178. (Data available at: http://godefroy.sdf-eu.org/apl95/ratios95.zip, 05.03.2013).

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

Vitanyi, P., 2005. Universal similarity. In:*ITW2005*, Rotorua, New Zealand.

Xu, R. and Wunch,D.C., 2009. Clustering. *John Wiley & Sons*, pp. 263-278.