GEOLOGIJA. 2006. Vol. 55. P. 48–57

Sedimentologija • Sedimentology • Седиментология

Clustering of descriptive-textual data on Silurian rocks of Lithuania

Valdas Rapševičius,

Algimantas Juozapavičius,

Antanas Brazauskas

Rapševičius V. , Juozapavičius A., Brazauskas A. Clustering of descriptive-textual data on Silurian rocks of Lithuania. *Geologija.* Vilnius. 2006. Vol. 55. P. 49–58. ISSN 1392-110X

Authors present a new approach to the clustering of descriptive-textual geological data with the aim to identify geological objects. Data used for the research come from the description of boreholes drilled in Lithuania and representing the Baltic Silurian basin. The problems under study include data preparation for analysis, methods of evaluating the number of clusters for Silurian data clustering, attribute elimination, techniques for data clustering and uncertainty elimination. A comprehensive description of clustering results is presented by an expert of geology.

The proposed method of clustering is of general purpose and can be applied to data that come also from any other domain alongside geology. Clustering results and their interpretation presented in the paper can be used for the Lithuanian Silurian overview only, because the identified rock types (clusters) strongly depend on the properties of the primary dataset. Those properties include subjective nature, multiple authorship and the unequal level of detailing.

Key words: Baltic Silurian basin, facial zones, descriptive-textual geological data, clustering, K-mode algorithm

Received 7 February 2006, accepted 20 March 2006

Valdas Rapševičius, Algimantas Juozapavičius, Faculty of Mathematics and Informatics, Vilnius University, Lithuania. E-mail: v.rapsevicius@it.lt, algimantas.juozapavicius@maf.vu.lt, Antanas Brazauskas, Faculty of Natural Sciences, Vilnius University, Lithuania. E-mail: antanas.brazauskas@gf.vu.lt

INTRODUCTION

Classical geology, for a long time being a rather qualitative than a quantitative science, is increasing its pace of using methods of data mining. Data mining algorithms are valuable for their ability to extract the knowledge from statistical properties of data and show an explosive growth in many applications (Thuraisingham, 1999), as well as in geology. The spread of computers was the primary reason of explosive growth of data analysis methods and computing tools in various topics of Earth sciences (Davis, 2002). Nevertheless, geologists till now often base their analysis on original data, not processed or mined. In many cases these data are not as much computer friendly as analysts would like to have. Most of historically collected borehole data in Lithuania are represented as the large scale descriptive-textual databases (Geolis, 1991–2005; Litosfera, 1997–2005), and in many cases these databases do not fit to any data analysis method forthright. Even more, data cannot be easily converted to a numeric or other precise format, therefore the sophisticated structuring methods are needed. This article presents such a new processing and structuring method, opening the way to cluster initially textual-descriptive data, later categorical data with an application to geological data.

INITIAL DATASET OF GEOLOGICAL DATA

The geological data used in this research were collected after drillings performed in Lithuania in the middle of the 20th century in search of oil fields. Many volumes of reports from those expensive drillings are now stored at the Geological Survey of Lithuania (www.lgt.lt).

Most of their texts are written in Russian as unstructured texts. Starting from 1997, most of these volumes of information have been partially put into MS Access tables (in the framework of the state "Litosfera" program). During this input, the process data were partially structured, namely the attributes of rock title, color, texture, structure and others were indicated. The primary dataset consists of 54 boreholes, 4030 layers overall (Litosfera, 1997–2005). Each layer is de-scribed with 35 lithological and palaentological parameters for primary and secondary rocks separately, plus 9 general parameters for the entire layer (e. g., oil/gas collector description, other rocks, contacts between rocks, upper contact, etc.). These data describe Silurian layers of Lithuania, the SE slope of the Silurian Baltic basin. Silurian system is one of the most complete, thick and best explored ones (Lithuanian Geology, 1994, p. 68) in the Lithuanian cross-section. The amount of Silurian data gives a possibility, on the one hand, to explore the layers in detail and on the other hand show that special methods of data analysis are desperately needed. Till now those tables are used only as electronic dictionaries to support decisions and analyses performed with the use of different data, e. g., palaeontological, geochemical, geophysical and others.

CATEGORICAL DATA

As defined in (Huang, 1997), let A_1, A_2, …, A_m be m attributes describing the space W and the $DOM(A_1)$, $DOM(A_2)$, …, $DOM(A_m)$ domains of these attributes. The domain $DOM(A_j)$ is defined as categorical if it is finite and unordered, e. g., for any a, $b \in DOM(A_j)$, either $a = b$ or $a \neq b$. A_j is called a categorical attribute. W is a categorical space, if all A_1, A_2, …, A_m are categorical. A categorical domain defined here contains only singletons. Though in the primary dataset we have combinational values, just like in (Gowda, 1991), later such values are not allowed. The same holds for a special value denoted in (Huang, 1997) by ε, i. e. defined in all categorical domains, and represents the missing values; it is allowed in the primary dataset, but later it is not defined.

Like in (Gowda, 1991), the categorical object $X \in W$ is logically represented as a conjunction of attribute-value pairs $(A_1 = x_1) \cap (A_2 = x_2) \cap \ldots \cap (A_m = x_m)$, where $x_j \in DOM(A_j)$ for $j=1\ldots m$. In (Michalski, 1983) an attribute-value pair $(A_j = x_j)$ is called a selector. Like in (Huang, 1997), we consider that X is a vector $(x_1, x_2, \ldots, x_m)$ and every object in W has exactly m attribute values.

Here and below in this article, like in (Davis, 2002, p. 7), we introduce a binary-state domain that falls into categorical attribute concept and indicates the presence or absence of lithological, palaentological or other feature. Formally, a binary-state attribute is the categorical attribute that has a domain containing flag value: $DOM(A_j) = (1,0)$. Domain values can be defined as *TRUE-FALSE*, *YES-NO*, *ON-OFF*, and so on. Here and below we use $(1,0)$ for binary-state domain values.

K-modes algorithm

K-modes algorithm (Huang, 1997) is a modification of the well-known k-means clustering algorithm (MacQueen, 1967; Anderberg, 1973), to cluster categorical data. Formally, both clustering methods are based on minimization of the functional

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} y_{i,l} d(X_i, Q_l), \quad (1)$$

where d is a dissimilarity measure, e.g. Euclidean distance in k-means algorithm. As defined above, the categorical at,tribute has a much more limited set of operations, therefore the dissimilarity measure, like in (Huang, 1997), could be defined as

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j), \quad (2)$$

where

$$\delta(x_j, y_j) = \begin{cases} 0, (x_j = y_j), \\ 1, (x_j \neq y_j). \end{cases} \quad (3)$$

The mode of the set X is the vector $Q=(q_1, q_2, \ldots, q_m)$ in categorical space W which minimizes the functional

$$D(Q, X) = \sum_{i=1}^{n} d(X_i, Q), \quad (4)$$

where $X=\{X_1, X_2, \ldots, X_n\}$, d can be defined as (2). Q is not necessarily an element of X and is composed of the most frequent values (modes) of attributes in X.

Here we define the steps of the k-modes algorithm, similarly as introduced in (Huang, 1997):

1) select k initial modes, one for each cluster,

2) allocate an object to the cluster whose mode is the nearest to it according to d (2). Update the mode of the cluster after each allocation by minimizing (4),

3) after all objects have been allocated to clusters, retest the dissimilarity of objects against the current modes. If an object is found such that its nearest mode belongs to another cluster rather than its current one, reallocate the object to that cluster and update the modes of both clusters,

4) repeat 3 until no object has changed clusters after a full cycle test of the whole data set.

Like in k-means algorithm, the k-modes one also produces locally the optimal solution which is dependent on the initial modes and the order of objects in the data set.

K-modes for binary-state values

The described k-modes clustering method fully meets the needs of binary-state domain attributes, because those

are categorical as well. But we would like to simplify clustering operations by following the nature of the binary-state domain. The dissimilarity measure could be defined as

$$d(X,Y)=\sum_{j=1}^{m}(x_j-y_j)^2. \quad (5)$$

Like in (Pedrycz, 2005, p.3), in the Hamming distance the square could be changed to absolute value brackets; this kind of k-means variation by (Bradley, 1997; Fung, 2001, p.11) is called k-median. The mode of the set can be found by exceeding the threshold of the related frequency:

$$q_i(X)=\begin{cases}1, (\frac{1}{n}\sum_{j=1}^{n}x_{i,j}\geq T)\\ 0, (\frac{1}{n}\sum_{j=1}^{n}x_{i,j}<T),\end{cases} \quad (6)$$

where *T* is a threshold of related frequency (in our example we have used *T=0.5*), and it is being compared with the relative frequency of the attribute.

Geological Dataset Preparation for Analysis

As stated above, the initial database of Lithuanian Silurian lithological data has a descriptive-textual unstructured nature and has to be prepared for clustering. A brief example of initial Silurian data of rock description would be "clayey limestone" for "rock title", "grey with greenish shade" for "rock color" and so on; 35 textual descriptive attributes in total. Eleven attributes that describe separately the primary and secondary rock of the layer have been chosen for further analysis. In this article, we will concentrate on individual rock samples and will not pay attention either it was initially taken from the primary or from the secondary rock of the layer and in what part of the country and how deep the drillings where made.

Formally, a record in the database represents a rock *R*. *R* is defined by or is composed of a finite number *m* of attributes *A* (in our case *m = 11*) $R=\{A_1,A_2,...,A_m\}$. Each attribute A_j has a finite domain of attribute values *DOM(*A_j*)* which consists of k_j terms $DOM(A_j)=\{t_1^j,t_2^j,\ldots,t_{kj}^j\}$.

The initial dataset should be transformed in this way: each rock *R* is to be represented by $R=\{t_1^1,t_2^1,\ldots,t_{k1}^1,t_1^2,t_2^2,\ldots t_{k2}^2,\ldots,t_{km}^m\}$, i. e. all the terms from all attribute domains now become individual attributes. The newly constructed attributes are represented by the binary-state domain $DOM(t_i^j)=\{0,1\}$, and the entire dataset has $\sum_{j=1}^{m}k_j$ attributes (our experimental Silurian transformed dataset consists of 205 attributes). The *R* attribute t_i^j value equals *1*, if the term t_i^j appears in A_j of *R*, and if t_i^j equals *0* – the feature t_i^j is absent in A_j of *R*. The new attributes have been named by concatenating the name of the initial attribute A_j, dollar sign ($) and the term t_i^j, e. g. "title$limestone", "title$clayey", "color$grey", "color$greenish" and so on.

This newly created dataset has exploded in dimensionality but, on the other hand, it has drastically reduced the complexity of attribute domains. Later in this paper we will show the technique of dimensionality reduction, which is very common to data mining.

Number of clusters

Initially we did not keep on any of available Baltic Silurian basin models (Нестор, Эйнасто 1977; Brazauskas, 1993), so it is said that the number of clusters, i. e. rock classes, is not known. We follow the non-parametric clustering approach as in (Fung, 2001, p. 14) where the number of clusters or groups can be determined as a function of some merging threshold. To describe the method of finding the number of clusters, we have to introduce two concepts of sums of dissimilarities, which are used in many data analyses, especially in clustering applications (Barr, 2004; Pribyl, Jain 1997; Zvika, 2004) that are also related to some other case grouping methods. We introduce local (or like in (Barr 2004) within-cluster) and global dissimilarities: local dissimilarity is a sum of dissimilarities between the cluster mode and observations inside a cluster; global dissimilarity is a sum of all cluster local dissimilarities. In the best case, the global dissimilarity measure would represent a global optimum of the dataset for a given number of clusters.

Here we have to note that as initial modes *k* of *k* clusters, *k* distinct random records from the dataset have been chosen. As (Huang, 1997; Fung, 2001, p.11) and many others emphasize, k-means as a technique behavior and the final *k* cluster centers strongly depend on the initial ones. So, by applying the modified k-modes algorithm to a dataset we have produced a number of results. Those are shown in Fig. 1: the upper line shows global dissimilarity measure and the lower one indicates the average local dissimilarity measure calculated by dividing the upper line by the number of clusters. The calculated power trends are very close to the actual values: R^2 equals 0.965 for global and 0.999 for average local.

It is trivial to note that a local dissimilarity varies from the sum of dissimilarities of all observations to the mode of the set (*k = 1*) to 0, where *k* equals to distinct attribute combinations of the dataset (in the worst case *k = n*). This kind of approach is met in hierarchical clustering, e. g. (Fung, 2001, p.14; Rapševičius, 2001) where the number of clusters could be determined by a threshold: if a threshold equals zero, the number of clusters is equal to the number of data points,

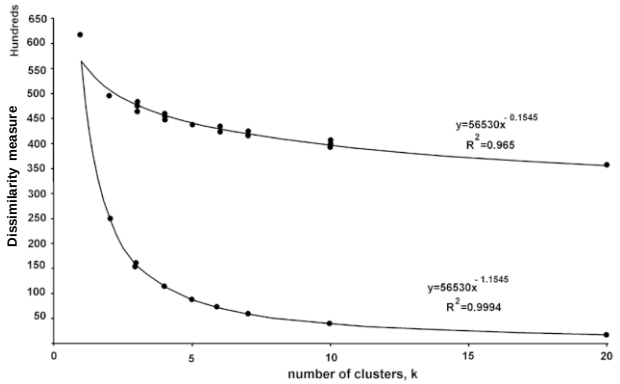

Fig. 1. Global and average local dissimilarity measures on different number of clusters

1 pav. Globalus ir vidutinis lokalus skirtybės matas, gautas grupuojant stebėjimus su nuolat didinamu klasterių skaičiumi

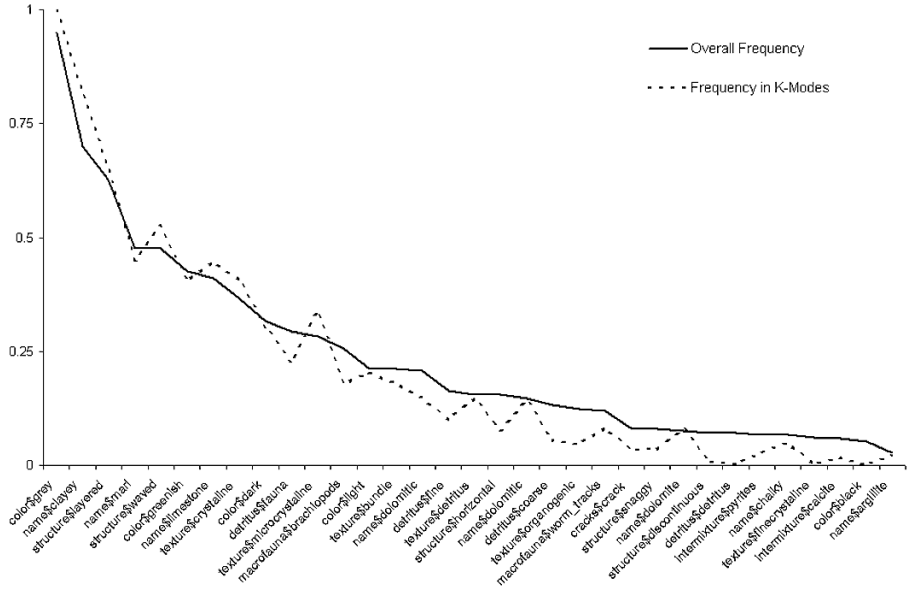

Fig. 2. Graph displays the relative frequencies of attributes that have been chosen for the further analysis. At least one of those attribute frequencies have exceeded 0.05 (5%)

2 pav. Santykiniai pasirinktų atributų dažnių grafikai: pradinių duomenų aibėje (ištisinė linija) ir k-modose (nutrūkstanti linija). Atributai atrinkti su sąlyga, kad bent vienas iš dažnio grafikų viršija 0,05 (5%) ribą

and with a high threshold the data are partitioned in just one single cluster. Thus, from the average local dissimilarity trend equation (the least squares fit through points by using the following equation: $y = cx^b$, where c and b are constants) we have found that the average local dissimilarity would reach 95% of its total amplitude (from k = 1, to k equal distinct attribute combinations of the dataset), then the number of clusters would be 14. Reliability limit of 5% was the value we have a priori agreed on.

Dimensionality reduction

After the number of clusters is found, we have to reduce the dimensionality of the dataset. Though there are a lot of methods in data mining how to reduce dimensions (Ye, 2003, p. 324; Hand, 2001, p. 118) we have rejected the "loosy" methods that produce new artificial dimensions, like (Hand, 2001, p. 120): factor analysis, projection pursuit and others. The idea was not to loose the meaning of an attribute to be able to use it straight for an interpretation. Reducing the dimensionality of data by deleting unsuitable attributes improves the performance of the clustering algorithm. More importantly, dimensionality reduction yields a more compact, more easily interpretable representation of the target concept, focusing the geologist's attention on the most relevant variables.

The combination of relative frequencies of observations is a method chosen in this research to eliminate dimensions. The combination includes 1) overall relative frequency of attribute values and 2) relative frequency of attribute in modes of k (14) clusters. In Fig. 2 we have shown the described frequencies of a reduced attribute set, except the first two attributes (color$grey and name$clayey). Such attributes have been removed because of a very high frequency. At least one of the chosen attribute frequencies has exceeded 0.05 (5%). The whole set of categorical attributes is shown in Fig. 3, and the name$argilitus has been included because of its importance to the domain.

Resolving uncertainties

After clustering the reduced dataset we have found that a large number of records (1833 records, 24.5%) became uncertain – they had the same smallest dissimilarity measure (4) with more than one cluster. To solve the uncertainties, we have introduced cluster modes with attributes having different importance that come as a final record partitioning step in the clustering process after the cluster modes have already been found (the overall clustering process is shown in Fig. 4).

After the final modes $Q=(q_1, q_2, \ldots, q_m)$, where $DOM(q_j) = (1,0)$ for each cluster have been found, let us change the mode attribute domain into $DOM(q_j) = \Im$ for each $q_j=1$ and define $q_j = \Theta$ for each $q_j=1$, where Θ is an integer constant greater than or equal to 1. Let us name $q_j>0$ a weighted and $q_j=0$ a not weighted mode attribute. For this particular partitioning process, let's define the new look of (3):

$$\delta(x_j, y_j) = \begin{cases} 0, x_j = y_j \cup (x_j > 0 \cap y_j > 0) \\ 1, (x_j \neq y_j \cap (x_j = 0 \cup y_j = 0)) \end{cases}. \quad (7)$$

Attribute	Overall dataset		Modes of k=14	
	Sum	R. freq.	Sum	R. freq.
structure$layered	4547	0.59	97	0.69
name$marl	3402	0.44	61	0.44
structure$waved	3393	0.44	79	0.56
color$greenish	3094	0.40	57	0.41
name$limestone	3044	0.39	60	0.43
texture$crystaline	2654	0.34	55	0.39
color$dark	2491	0.32	48	0.34
detritus$fauna	2191	0.28	31	0.22
texture$microcrystaline	2135	0.28	51	0.36
macrofauna$brachiopods	1974	0.25	30	0.21
color$light	1773	0.23	34	0.24
texture$bundle	1637	0.21	33	0.24
name$dolomitic	1329	0.17	13	0.09
detritus$fine	1291	0.16	20	0.14
texture$detritus	1222	0.16	24	0.17
structure$horizontal	1162	0.15	11	0.08
name$dolomitic	1115	0.14	10	0.07
detritus$coarse	1038	0.13	14	0.10
texture$organogenic	982	0.12	18	0.13
macrofauna$worm_tracks	902	0.11	11	0.08
cracks$crack	629	0.08	5	0.04
structure$snaggy	624	0.08	3	0.02
name$dolomite	529	0.06	12	0.09
structure$discontinuous	506	0.06	1	0.01
detritus$detritus	479	0.06	2	0.01
intermixture$pyrites	457	0.06	0	0.00
name$chalky	445	0.05	1	0.01
texture$finecrystaline	417	0.05	3	0.02
intermixture$calcite	410	0.05	5	0.04
color$black	403	0.05	3	0.02
name$argillite	224	0.02	4	0.03

Fig. 3. Reduced attribute set
3 pav. Analizei atrinkti atributai

The (7) produces exactly the same result as like (3), but takes into account the weighted binary-state domain of the mode. Now let us introduce a similarity measure *s*,

$$s(X,Y)=\sum_{j=1}^{m}\sigma(x_j,y_j), \tag{8}$$

where

$$\sigma(x_j,y_j)=\begin{cases}x_j\times y_j,(x_j>0\cap y_j>0)\\1,(x_j=y_j=0)\\0,(x_j\neq y_j\cap(x_j=0\cup y_j=0))\end{cases}. \tag{9}$$

As (Zvika, 2004) has noted, it is trivial that if mode attributes would be defined via $DOM(q_j) = (1,0)$, it would meet the equality

$$m = s(X,Y) + d(X,Y), \tag{10}$$

where m equals to the number of dimensions in a dataset. Both similarity and dissimilarity measures in (Hand, 2001, p. 25) are called proximity measures. For the weighted mode attributes $DOM(q_j) = \Im$ we meet the inequality

$$s(X,Y) \geq m - d(X,Y). \tag{11}$$

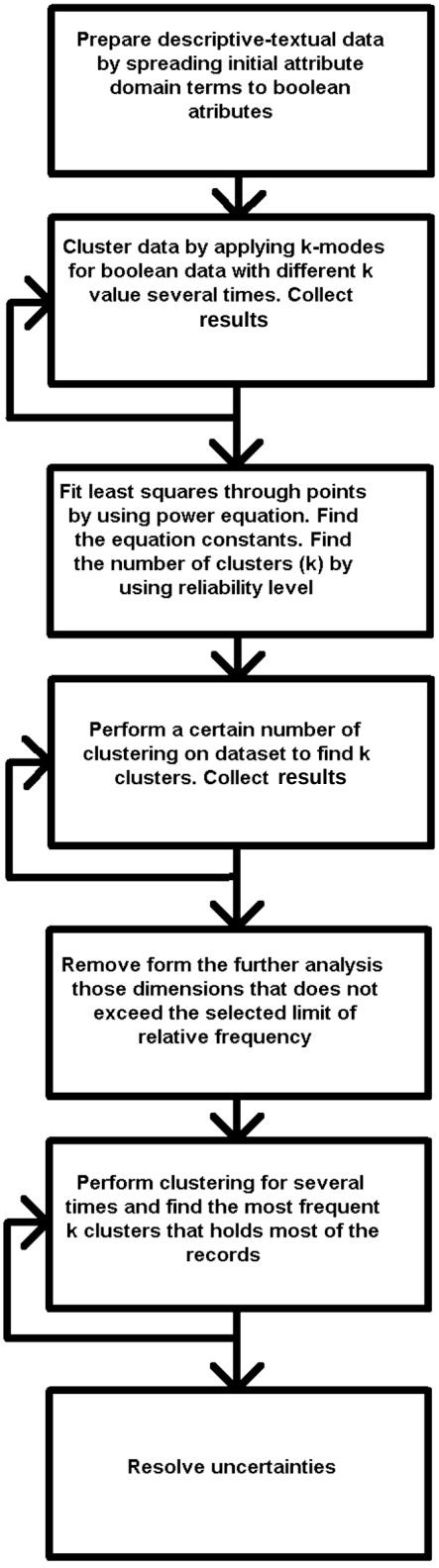

Fig. 4. Overall clustering process
4 pav. Visas stebėjimų grupavimo (klasterizacijos) procesas

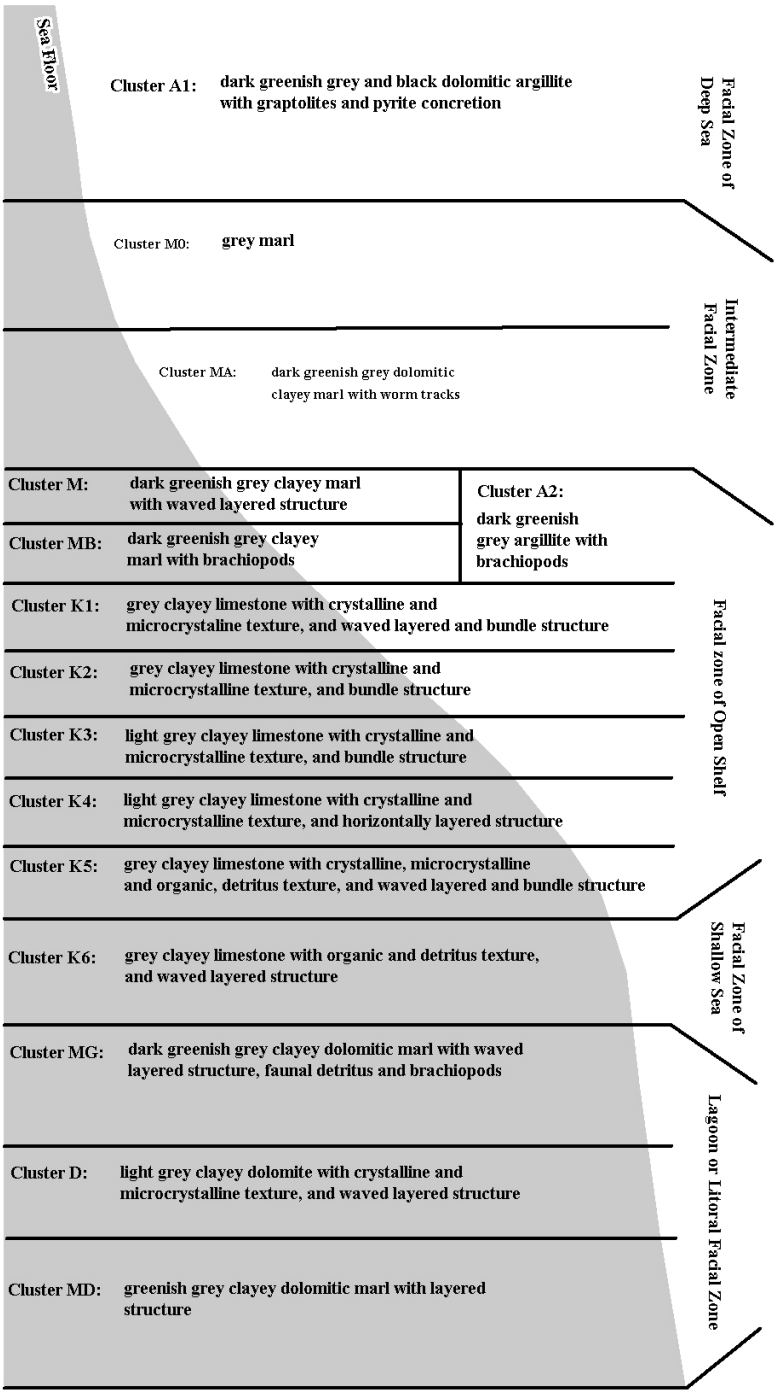

Fig. 5. Clusters and facial zones of (Нестор, Эйнасто, 1977)
5 pav. Proceso metu išskirtų uolienų grupių (klasterių) ir facijų zonų (Нестор, Эйнасто, 1977) sugretinimas

The above approach of combined similarity and dissimilarity measures, first, raises the importance of the observed (1) terms against unobserved (0) and, secondly, allows an expert to define individual attribute weights in different modes. The algorithm to resolve uncertainties would be the following:

1) $q_j = q_j + 1$ for each $q_j > 0$,

2) partition records by minimizing dissimilarity measure (2) and in the case of uncertainty – maximize similarity measure (8),

3) repeat from step 1 if the number of unresolved uncertainties has changed,

4) reduce Θ to the smallest stable weight constant $q_j = q_j - 1$ for each $q_j > 0$,

5) construct a matrix $m \times m$ and put into its cells the number of uncertainties between the attributes,

6) review the matrix and increase the weight of mode attributes of the conflicting mode pair that are weighted ($q_j > 0$) in one conflicting mode and are not weighted ($q_j = 0$) in another,

7) repeat partitioning by minimizing the dissimilarity measure (2) and, in the case of uncertainty, maximizing the similarity measure (8),

8) if the number of uncertainties left does not satisfy the limit of reliability – continue from step 4.

After applying steps 1 to 3 twice ($\Theta = 2$), the number of uncertainties has decreased to 772 records (10.3% of the total) and became stable. By applying steps 4 to 7 twice we arrived at 264 (3.46%) uncertainties, which are far below the limit of 5%.

EXPERIMENTAL RESULTS

The overall clustering process is shown in Fig. 4, and we hope that it gives the reader the idea of clustering large descriptive-textual geological databases. Below we provide the description of final clusters (after geologist expertise we have defined 15 instead of 14 clusters) by referring to the (Нестор, Эйнасто 1977) model of the Baltic Silurian basin. All 15 clusters have been labeled to be easier recognizable by the human. The clusters and facial zones are presented in Fig. 5.

Cluster A1: dark greenish grey and black dolomitic argillite with graptolites and pyrite concretions. This class of rock belongs to the deepest part of the Baltic Silurian basin: facial zone V in (Нестор, Эйнасто, 1977). A1 cluster has been assigned to the deepest facial zone of the Baltic Silurian Basin, because it has almost no macro fauna such as brachiopods or corals; quartz and mica are more frequent than in other clusters, and the reaction with HCl is very poor.

Cluster M0: grey marl. Into this cluster have fallen all the “marls” that have a very poor description, thus

in general this rock class could be attached to any part of the basin where marls are met. But we have attached it to the lower part of the Intermediate Facial Zone (IV_2), based on the fact that there significant amounts of graptolites, organic matter, and pyrite concretions are observed. On the contrary, there are almost no benthic macrofauna.

Cluster MA: dark greenish grey dolomitic clayey marl with worm tracks. This kind of rock comes from the upper part of Intermediate Facial Zone (IV_1), because, just like in the above cluster, graptolites are still there. The amount of macrofauna (brachiopods and trilobites) constantly increases. All those facts show that the rocks of this cluster were formed in a slightly shallower depositional environment than the rocks of **M0** cluster. The macrofauna, except worm tracks and graptolites, is very rare. As follows from the above statements, we think it is reasonable to assign the **MA** cluster to the fourth facial zone.

Cluster A2: dark greenish grey argillite with brachiopods. This argillite comes from a far shallower zone than A1 and presents a regressive sequence in the bottom of Open-Shelf Facial Zone, III_3 litho-facies. Large amounts of micro-grained terrigeneous material were brought into the Baltic Silurian depositional basin after the regressive regime had settled down, especially in the last stages of the basin evolution cycle. After diagenesis those sediments transformed into argillite beds. However, the depositional basin was much shallower than the one that formed **A1** argillites as brachiopod fauna are much more frequent there. All other lithological and faunal features described in (Нестор, Эйнасто, 1977) are common to IV faunal zone.

The transgressive sequence in this point would be presented by the pack of two clusters: the lower **Cluster M** (dark greenish grey clayey marl with a waved layered structure) and the upper **Cluster MB** (dark greenish grey clayey marl with brachiopods). Waved layered rock structures are formed in the facies of marl with the bundles of clayey limestone that are common to the Open Shelf facial zone. This facies is characterized by frequent scour surfaces, detritus of brachiopods and crinoids, worm tracks and threads of argillites. All the above features are common to the rocks that come from **M** cluster. However, **MB** cluster rocks have a much wider variety and larger amount of macro fauna.

Cluster K1: grey clayey limestone with a crystalline and microcrystalline texture and a waved layered and bundle structure. Rocks like this are met in the lower part of III_2 litho-facies of the Open-Shelf Facial Zone. Gray and light gray limestone of various structure and texture with lots of detritus is natural in this facial zone. Yet gray colors are more natural to the lower (deeper part of the basin) and light ones to the upper part of the zone. Most of the rocks in the lower part of the third zone are gray, microcrystalline, waved layered and bundle structured.

Cluster K2: grey clayey limestone with a crystalline and fine crystalline texture and layered structure. This rock is common to the lower part of III_2 litho-facies of the Open-Shelf Facial Zone. The middle- to coarse crystalline structure of K2 cluster rocks and intense reaction with HCl acid are the main features that differentiate this cluster from **K1**. Those rocks have a relatively high content of marl and detritus of graptolites fauna, thus showing that K2 rocks come from the lower part of III_2 litho-facies.

Cluster K3: light grey clayey limestone with a crystalline and microcrystalline texture and bundle structure. It is common to the mid of III_2 litho-facies of the Open-Shelf Facial Zone. Rocks of the upper part of the III facies zone have various structures: waved and bundle layered, bundle, etc. Gradually organogenic and detritus textures like crinoidae, coral and stromatopore appear. The macrofauna is diverse yet not numerous; squids are predominant. All the above features and faunal singularity are the main criteria to ascribe **K3** cluster to the middle part of III_2 litho-facies of the Open-Shelf Facial Zone.

Cluster K4: light grey clayey limestone with crystalline and microcrystalline texture, and horizontally layered structure. This rock is common to the lower part of III_1 litho-facies of the Open-Shelf Facial Zone. The main features that distinguish this cluster from the others include brachiopodal and crinoidae texture and a small amount of clayey material. The horizontally layered structure and white rock color could be mentioned as well.

Cluster K5: grey clayey limestone with a crystalline, microcrystalline and organic, detritus texture and waved layered and bundle structure. This rock is common to the upper part of III_1 litho-facies of the Open-Shelf Facial Zone. The sedimentary environment of the rocks of this cluster could be described as before-reef (towards the open sea). Textural features like organogenic, detritus, etc. show that. Other important indicators of reef proximity would be a bent layered structure, calcite concretions and sockets, and many broken fragments of other rocks.

Cluster K6: grey clayey limestone with organic and detritus texture, and waved layered structure. This kind of rock comes from the Shallow-Sea Facial Zone (II) and is related to reefs. This kind of rocks usually comes from sedimentary domains affected by a strong hydrodynamic influence. Organigenic and broken organigenic textures are the main features that distinguish this cluster from the other limestone rock clusters. The rocks of this cluster contain a lot of reef-forming macrofauna – stromatoporoids, corals, crinoidea. Towards the open sea this cluster switches to light clayey limestone (cluster **K5**), while towards the coast it turns into dark gray dolomitic marl (cluster **MD**).

Cluster MG: dark greenish grey clayey dolomitic marl with a waved layered structure, faunal detritus and brachiopods. Together with MD and D clusters this rock forms the shallowest part of the basin; the Lagoon-Littoral

Facial Zone (I). Dark grey and even black colors show the slow-motion circulation of water in the sedimentary basin where the rocks of this cluster were formed. Such a condition is common to closed or semi-closed lagoons between the reef and the coast. Large amounts of brachiopods and ligulidae, yet not diverse in species, are quite often met in the rocks like this.

Cluster D: light grey clayey dolomite with a crystalline and microcrystalline texture and waved layered structure. This rock class would be ascribed to one of the shallowest parts of the basin, the mid of the Lagoon-Littoral Facial Zone (I). The porous, cavernous structure, exceptionally lingulid macrofauna, and yellow and red color are the main features that allow us to assign this cluster to the above Facial Zone.

Cluster MD: greenish grey clayey dolomitic marl with a layered structure. This class rock represents the shallowest part of the basin, the upper part of the Lagoon--Littoral Facial Zone (I). Rocks of this facial zone have a red and purple color and spotted structure, include concretions of gypsum and limonite. Those rocks contain almost no macrofauna, except a few specific fish species.

Future tasks

Though there are a lot of possible future workarounds in the domain of data and tools presented in this paper, we suggest two areas of primary interest. One area is related to the spatial nature of data: each rock sample could be presented as a point in a 3D space. It is obvious that lithological attributes have different importance depending on the location. It is possible to have attributes allowing elimination, because they are rare in the dataset as a whole, but they could be very important in particular subsets of data. Another interesting issue to be solved is related to the sequences of rocks. Sequences in geology are very important and present transgress, regress or a mixed state of a sedimentary basin. We hope that clustering and sequence analysis techniques could bring a fresh insight into the history of the Baltic Silurian basin.

Tools used

The Original Silurian data were entered into Oracle 9i database. All the data preparation and transformation operations where conducted in Oracle database by using PL/SQL programming language. As a result, a package of applied procedures and functions has been compiled. Clustering and other operations have been programmed by using Perl.

CONCLUSIONS

Authors propose a process to cluster descriptive-textual data, which has proven to be efficient, scalable and could be fully or at least semi-automatic. Most of human attention and thus time has to be paid to the data preparation (actually the first) stage; all the other stages can be performed by the software, with the provided initial parameters. The method has identified geological objects (rock types) by analyzing a comparable huge amount of data; it was not possible before because of the unstructured character of data and their huge amount. The proposed method of clustering is of general purpose and can be applied also to data that come from any other domain alongside geology. The clustering results and their interpretation presented in this article can be used for the Lithuanian Silurian overview only, because the identified rock types (clusters) strongly depend on the properties of the primary dataset. Those properties include subjective nature, multiple authorship and the unequal level of detailing.

References

Anderberg M. R. 1973. Cluster Analysis for Applications. Academic Press, Inc., New York, NY.

Barr G., Dong W., Gilmore C. J. 2004. High-throughput powder diffraction. II. Applications of clustering methods and multivariate data analysis. *Journal of Applied Crystallography*. Vol. 37. Part 6. 874–882.

Davis J. C. 2002. Statistics and Data Analysis in Geology. John Wiley & Sons Inc. 638 p.

Fung G. 2001. A Comprehensive Overview of Basic Clustering Algorithms. Unpubl. 35 p.

Geolis. 1991–2005. Geological Borehole Register "Geolis", Geological Survey of Lithuania.

Gowda K. C., Diday E. 1991. Symbolic Clustering Using a New Dissimilarity Measure. *Pattern Recognition*. ***24(6)***. 567–578.

Hand D., Mannila H., Smyth P. 2001. Principles of Data Mining. The MIT Press. 546 p.

Huang Z. 1997. A fast clustering algorithm to cluster very large categorical data sets in data mining: *Proc. SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*. Tucson, Arizona, USA, May 1997. 146–151.

Jain A. K., Zongker D. 1997. Representation and recognition of handwritten digits using deformable templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. ***19(12)***.

Kantardzic M. 2003. Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons. 343 p.

Lithuanian Geology. 1994. Monograph. Vilnius: Mokslo ir enciklopedijų leidykla (in Lithuanian). 447 p.

Litosfera. 1997–2005. Geological Database "Litosfera". Faculty of Natural Sciences, Vilnius University, Geological Institute of Lithuania.

MacQueen J. B. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. 281–297.

Michalski R. S., Stepp R. E. 1983. Automated construction of classifications: conceptual clustering versus numerical taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. ***5(4)***. 396–410.

Bradley P. S., Mangasarian O. L., Street W. N. 1997. Clustering via concave minimization. M. C. Mozer, M. I. Jordan, T. Petsche (ed.). *Advances in Neural Information*

Processing Systems -9-, pages 368–374. Cambridge, MA, 1997. MIT Press. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/96-03.ps.

Paskevičius J. 1997. The Geology of the Baltic Republics. Vilnius. 113–116 p.

Pedrycz W. 2005. Knowledge-Based Clustering from Data to Information Granules. Wiley-Interscience. 2003. 315 p.

Pribyl O. Clustering of activity patterns using genetic algorithms. Unpubl. presentation in the Pennsylvania State University. 9 p.

Rapsevicius V., Juozapavicius A. 2001. Clustering through decision tree construction in geology. *Nonlinear Analysis: Modelling and Control*. ***6(2)***. 29–41.

Thuraisingham B. 1999. Data Mining. Technologies, Techniques, Tools, and Trends. CRC Press.

Trevor H., Tibshirani R., Friedman J. 2001. The Elements of Statistical Learning. Data Mining, Inference, and Prediction (Springer Series in Statistics), Springer.

www.lgt.lt, website of Lithuanian Geological Survey: http://www.lgt.lt

Ye N. (editor). 2003. The Handbook of Data Mining. Lawrence. 720 p.

Zvika M. 2004. Structure based computational aspects of similarity and analogy in natural language: unpubl. thesis in Hebrew University.

Бразаускас А. З. 1993. Конодонты и биостратиграфия силура Литвы. Докторская диссертация. Вильнюс. 336 с.

Нестор Х. Э., Эйнасто Р. Э. Фациально-седиментологическая модель силурийского Палеобалтийского периконтинентального бассейна. *Фации и фауна силура Прибалтики.* Таллин, 1977. 89–121 с.

Valdas Rapševičius, Algimantas Juozapavičius, Antanas Brazauskas

TEKSTINIŲ APRAŠOMŲJŲ LIETUVOS SILŪRO UOLIENŲ DUOMENŲ KLASTERIZACIJA

Santrauka

Šiame straipsnyje yra pateikiamas geologinių objektų klasterizacijos metodas naudojant tekstinius aprašomuosius duomenis. Metodą rekomenduojama taikyti tais atvejais, kai turima daug tekstinės aprašomosios, dažniausiais kokybinės informacijos, o kiekybinių objektyvių duomenų arba neįmanoma gauti, arba jie susiję su didelėmis laiko ir lėšų sąnaudomis. Metodui pagrįsti buvo naudojami Lietuvos silūro 54 gręžinių tekstiniai aprašomieji duomenys iš 4030 sluoksnių. Aprašytos kiekvieno sluoksnio pirminė ir antrinė uolienos, kiekvienai jų pateikti 35 litologiniai ir paleontologiniai parametrai. Visa šių duomenų aibė iki šiol nebuvo nagrinėta jokiais statistiniais metodais.

Tyrimo metu tekstiniai aprašomieji uolienų duomenys buvo skaidomi į elementariuosius derinius tol, kol jų skaidymas turėjo semantinę prasmę. Vėliau deriniai buvo išskleisti į atributus, kurių reikšmės įgyja 1, jei derinys randamas uolienoje, ir 0 – jei derinio uolienoje nėra.

Gautos matricos analizei buvo pasitelktas (Huang, 1997) aprašytas k-modų algoritmas, kuris yra gerai žinomo k vidurkių klasterizacijos algoritmo (MacQueen, 1967; Anderberg, 1973) modifikacija kategoriniams duomenims klasifikuoti.

Klasterių skaičius buvo nustatomas parametriniu metodu: modifikuotas k-modų algoritmas buvo skaičiuojamas įvairiam klasterių skaičiui – nuo 1 (visi geologiniai objektai priklauso vienai klasei) iki 20, kiekvienu atveju perskaičiuojant bendrą ir vidutinę duomenų aibės atstumų iki klasterių centrų sumą. Iš gautų duomenų sudarytas grafikas (1 pav.), kuriame tiek bendra, tiek vidutinė atstumų iki klasterių centrų suma kinta pagal jėgos (laipsnių) dėsnį. Remiantis pasirinkta 5% leistina klaidos tikimybe, nustatytas optimalus 14-os klasterių skaičius.

Siekiant sumažinti dimensijų (atributų) skaičių buvo skaičiuojamas kiekvieno duomenų aibės atributo ir 14-oje klasterių centrų pastebėtų atributų dažniai. Atlikus skaičiavimus buvo atmesti pernelyg dažni (spalva: pilka ir pavadinimas: molingas) mažesnes nei 0,5 (5%) abi dažnių reikšmes turintys požymiai. Tolimesnei analizei liko 30 požymių plius pavadinimas: argilitas, kuris yra svarbus dalykinis požymis.

Atlikus galutinės duomenų aibės klasterizaciją pastebėta, kad labai didelė jų dalis (1833 įrašų, 24,5%) įgyja vienodą minimalų skirtybės matą (atributų skirtumų sumą) su keletu klasterių centrų ir nėra vienareikšmiškai aišku, kuriai klasei reikėtų priskirti objektą. Straipsnyje detaliai aprašytas metodas, kuris išsprendžia šią problemą: „pasimetusių“ objektų skaičius sumažintas iki 264 (3,46%).

Klasifikacijos metu gauta 15 klasterių, kurie straipsnio pabaigoje detaliai interpretuojami Lietuvos silūro eksperto bei priskiriami Baltijos silūro baseino facijų zonoms (Нестор, Эйнасто, 1977).

Валдас Рапшявичюс, Альгимантас Юозапавичюс, Антанас Бразаускас

КЛАСТЕРИЗАЦИЯ ТЕКСТОВЫХ ОПИСАТЕЛЬНЫХ ДАННЫХ СИЛУРИЙСКИХ ПОРОД ЛИТВЫ

Резюме

В статье представлен метод кластеризации геологических объектов с использованием текстовых описательных данных. Метод рекомендуется применять в тех случаях, когда имеются большие объемы информации качественного характера, а получение объективных данных не возможно или связано с большими затратами времени и средств. Для обоснования метода использованы текстовые описательные данные силурийских пород из 54 скважин Литвы, описание которых представлено из 4030 слоев. В каждом слое описаны первостепенная и второстепенная породы и каждая из них охарактеризована 35 литологическими и палеонтологическими параметрами. Вся эта совокупность данных до сих пор не была обработана никакими статис-тическими методами.

Во время исследования текстовые описательные данные пород расчленялись на элементарные комбинации до тех пор, пока их расчленение имело семантический смысл. Позже эти комбинации были развернуты на атрибуты с определенными значениями. Если комбинация встречается в породе, то она приобретает 1, и 0 – если комбинация не обнаруживается.

Для анализа полученной матрицы применен алгоритм к-мода (Huang, 1997), который является модификацией хорошо известного алгоритма (MacQueen, 1967; Anderberg, 1973) кластеризации по к-средних для классификации категорических данных.

Для определения числа кластеров применен параметрический метод, при котором модифицированный к-моды алгоритм вычислялся для разного числа кластеров – от 1 (все геологические объекты представляют один класс) до 20, в каждом случае вычисляя сумму общего и среднего расстояний совокупностей данных до центров кластеров. По полученным данным составлен график (рис. 1), из которого видно, что сумма средних расстояний до центров кластером меняется по закону степеней. Опираясь на допускаемом 5% уровне значимости, выделено оптимальное число 14 кластеров.

Чтобы уменьшить число атрибутов, вычислялись частоты атрибута в каждой совокупности данных и замеченных атрибутов в центрах 14 кластеров. После выполненных подсчетов из дальнейших вычислений были исключены очень часто повторяющийся (цвет: серый и название: глинистый), а также меньше чем 0,5 (5%) обеими значениями частот обладающие признаки.

После выполнения окончательной кластеризации совокупностей данных замечено, что большая часть записей – 1833 (24,5%) приобретают одинаковую минимальную меру различия (сумму различия атрибутов) с центрами некоторых кластеров, поэтому нет однозначности, к которому классу объектов следовало бы их присоединить. В статье детально описан метод, который решает эту проблему. С помощью этого метода число „растерявшихся“ объектов сократилось до 264 (3,46%).

В ходе кластеризации получены 15 кластеров и детально интерпретированы по фациальным зонам Балтийского силурийского бассейна согласно модели Х. Нестора и Р. Эйнасто (1977).