CLUSTERING WITH GRANULAR INFORMATION PROCESSING

Urszula Kużelewska
Faculty of Computer Science, Technical University of Bialystok, Wiejska 45a, 15-521 Bialystok, Poland

Keywords: Knowledge discovery, Data mining, Information granulation, Granular computing, Clustering, Hyperboxes.

Abstract: Clustering is a part of data mining domain. Its task is to detect groups of similar objects on the basis of established similarity criterion. Granular computing (GrC) includes methods from various areas with the aim to support human with better understanding analyzed problem and generated results. Granular computing techniques create and/or process data portions named as granules identified with regard to similar description, functionality or behavior. Interesting characteristic of granular computation is offer of multi-perspective view of data depending on required resolution level. Data granules identified on different levels of resolution form a hierarchical structure expressing relations between objects of data.
A method proposed in this article performs creation data granules by clustering data in form of hyperboxes. The results are compared with clustering of point-type data with regard to complexity, quality and interpretability.

1 INTRODUCTION

Granular computing (GrC) is a new multidisciplinary theory rapidly developed in recent years. The most common definitions of GrC (Yao, 2006), (Zadeh, 2001) include a postulate of computing with information granules, that is collections of objects, that exhibit similarity in terms of their properties or functional appearance. Although the term is new, the ideas and concepts of GrC have been used in many fields under different names: information hiding in programming, granularity in artificial intelligence, divide and conquer in theoretical computer science, interval computing, cluster analysis, fuzzy and rough set theories, neutrosophic computing, quotient space theory, belief functions, machine learning, databases, and many others. According to more universal definition, granular computing may be considered as a label of a new field of multi-disciplinary study, dealing with theories, methodologies, techniques and tools that make use of granules in the process of problem solving (Yao, 2006).

Distinguishable aspect of GrC is a multi-perspective standpoint of data. Multi-perspective means diverse levels of resolution depending on saliency features or grade of details of studied problem. Data granules that are identified on different levels of resolution form a hierarchical structure expressing relations between objects of data. Such structure can be used to facilitate investigation and helps to understand complex systems. Understanding of analyzed problem and attained results are main aspects of human-oriented systems. There are also definitions of granular computing additionally concentrating on systems supporting human beings (Bargiela and Pedrycz, 2002)-(Bargiela and Pedrycz, 2006). According to definitions mentioned above, such methodology can allow to ignore irrelevant details and concentrate on essential features of the systems to make them more understandable. In (Bargiela and Pedrycz, 2001) an approach of data granulation based on approximating data by multi-dimensional hyperboxes is presented. The hyperboxes represent data granules formed from the data points focusing on maximization of density of information present in the data. It benefits from improvement of computational performance among the others. The algorithm is described in the following sections.

Clustering is a part of data mining domain performing exploratory analysis of data. Its aim is to determine natural clusters, which means, groups of objects more similar to one another than to the objects from other clusters (A. K. Jain and Flynn, 1999). Criterion of similarity depends on clustering algorithm

Kużelewska U..
CLUSTERING WITH GRANULAR INFORMATION PROCESSING.
DOI: 10.5220/0003142700890097
In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence* (ICAART-2011), pages 89-97
ISBN: 978-989-8425-40-9

and data type. The most common similarity measure is distance between points, for example, Euclidean metric for continuous attributes. There is no universal method to assess clustering results. One of the approaches is to measure quality of partitioning by special indicants (validity indices). The most common measures are: Davies-Bouldin's (DB), Dunn's (Halkidi and Batistakis, 2001), Silhouette Index (SI) (Kaufman and Rousseeuw, 1990) and CDbw (Halkidi and Vazirgiannis, 2002). Clustering algorithms have wide applications in pattern recognition, image processing, statistical data analysis and knowledge discovery. Quoting definitions mentioned above, where granule is determined as a set of objects, one can consider groups identified by clustering algorithms as data granules. According to that definition, a granule can contain other granules as well as be the part of another granule. It makes possible to employ clustering algorithms to create granulation structures of data.

The article proposes an approach of information granulation by clustering data, that are in form of hyperboxes. Hyperboxes are created in the first step of the algorithm and then they are clustered by SOSIG (Stepaniuk and Kużelewska, 2008) method. This solution is effective with regard to time complexity and interpretability of generated groups of data. The paper is organized as follows: the next section, Section 2, describes proposed approach, Section 3 reports collected data sets as well as executed experiments. The last section concludes the article.

2 GRANULAR CLUSTERING BY SOSIG

The proposed method of data granulation is composed of two phases. First phase prepares data objects in form of granules (hperboxes), whereas second detects similar groups of the granules. The final result of granulation is a three-level structure, where the main granulation is defined by clusters of granules and the following level consists of granules from components of the top level cluster. The down third level consists of point-type objects.

The method of hyperboxes creation is designed to reduce the complexity of the description of real-world systems. The improved generality of information granules is attained through sacrificing some of the numerical precision of point-data (Bargiela and Pedrycz, 2001). The hyperboxes (referred as I) are multi-dimensional structures described by a pair of values a and b for every dimension. The point a_i represents minimal and b_i maximal value of the granule in i-th dimension, thus width of i-th dimensional edge equals $|b_i - a_i|$. Creation of hyperboxes is based on maximization of "information density" of granules (the algorithm is described in details in (Bargiela and Pedrycz, 2006)). Information density can be expressed by Equation 1.

$$\sigma = \frac{card(I)}{\phi(width(I))} \tag{1}$$

Maximization of σ is a problem of balancing the possible shortest dimensions against the greatest cardinality of formed granule I. In presented experiments in the following section, cardinality of the granule I is considered as the number of point-type objects belonging to the granule. Belonging means that the values of point attributes are between or equal to the minimal and maximal values of the hyperbox attributes. For that reason there is necessity to re-calculate cardinality in every case of forming a new largest granule from combination of two granules. In multi-dimensional case of granules, as a function of hyperboxes width, is applied a function from Equation 2:

$$\phi(u) = \exp(K \cdot \max_i(u_i) - \min_i(u_j)), i, j = 1, \ldots, n \tag{2}$$

where $u = (u_1, u_2, \ldots, u_n)$ and $u_i = width([a_i, b_i])$ for $i, j = 1, \ldots, n$. The points a_i and b_i denote respectively minimal and maximal value in i-th dimension. The constant K originally equals 2, however in the experiments there were used different values of K given as a parameter. Computational complexity of this algorithm is $O(N^3)$. However, in every step of the method, the size of data is decreased by 1, what in practice significantly reduces the general complexity. The data granulation algorithm assumes processing hyperboxes as well as point-type data. To make it possible new data are characterized by $2 \cdot n$ values in comparison with original data. The first n attributes describe minimal, whereas the following n describe maximal values for every dimension. To assure topological "compatibility" point-type data and hyperboxes dimensionality of the data is doubled initially.

2.1 Self-Organizing System for Information Granulation

The SOSIG (**S**elf-**O**rganizing **S**ystem for **I**nformation **G**ranulation) algorithm is a system designed for detecting granules present in data. The granulation is performed by clustering and the clusters can be identified on the different level of resolution. The prototype of the algorithm is a method described in (Wierzchoń and Kużelewska, 2006). However, in SOSIG granulation property and application to cope with different attributes types was introduced. This follows

fundamental changes in its implementation. In the following description of the algorithm there are used new terms and symbols in contrary to the description from (Wierzchoń and Kużelewska, 2006) more compatible with GrC theory. SOSIG creates a network structure of connected objects (in general points, but in the presented solution hyperboxes) forming clusters. Organization of the system, including the points as well as the connections, is constructed on the basis of relationships between input data, without any external supervision. The structure points are representatives of input data, that is, an individual object from the structure stands for one or more object from input set. In effect of this, the number of representatives is much less than input data without lost of information. To have convenient and compact notation, let us assume input data are defined as an information system $IS = (U,A)$ (Pawlak, 1991), where $U = \{x_1,\ldots,x_n\}$ is a set of objects and $A = \{a_1,\ldots,a_k\}$ is a set of attributes. The result generated by SOSIG is also described by an information system $IS' = (Y, A \cup \{a_{gr}\})$, where the last attribute $a_{gr} : Y \to \{1,\ldots,nc\}$ denotes label of generated cluster and $card(Y) \leq card(U)$ and $\forall x \in U \exists y \in Y(\delta(x,y) < NR)$. The parameter NR in general defines region of objects interactions and is described later. The steps of the main (learning) part of the SOSIG are shown in Algorithm 1, whereas initial phase is separated into Algorithm 2. Further classification of new as well as training points can be performed using so-created network as it is presented in Algorithm 6.

The parameter NR (Neighborhood Radius) existing in above descriptions defines neighborhood of objects from IS'. It directly influences level of granulation of the input set. Initial value of NR is proportional to maximal of nearest neighbor distances in the input set (see Equation 3).

$$NR_{init} = \max(\{\min(\{\delta(x_i,x_j) : x_j \in U \ \&\ x_j \neq x_i\}) : x_i \in U\}) \tag{3}$$

The following values of NR are calculated from current state of the network (see Equation 4).

$$NR = rg \cdot \frac{\sum_{y_i \in Y} \min(\{\delta(y_i,y_j) : y_j \in Y \ \&\ y_j \neq y_i\})}{card(Y)} \tag{4}$$

where $rg \in [0,1]$ is a resolution of granulation parameter. Value of rg is proportional to the level of granulation, that is the top, first level of granulated structure is characterized by lower values of rg and the following levels (second, third and so on) are characterized by higher values of the parameter. The rg values are not fixed for according level of hierarchy, but related to individual set of data. However, there is a value of $rg = 0.5$ the most often appearing in empirical tests for the most separated and compact (termed "natural") clustering.

After initial phase (like normalization of data, calculation initial value of NR) iterated steps of the algorithm follow. First, the system objects are assessed.

Algorithm 1: Construction of information system with a set of representative objects.

Data:

- $IS = (U,A)$ - an information system, where $U = \{x_1,\ldots,x_n\}$ is a set of objects and $A = \{a_1,\ldots,a_k\}$ is a set of attributes,
- $\{\delta_a : a \in A\}$ - a set of distance function of the form $\delta_a : V_a \times V_a \to [0,\infty)$, where V_a is a set of values for attribute $a \in A$ and a global distance function $\delta : U \times U \to [0,\infty)$ defined by $\delta(x,y) = fusion(\delta_{a_1}(a_1(x),a_1(y)),\ldots,\delta_{a_k}(a_k(x),a_k(y)))$
- $size_{net} \in \{0,1,\ldots,card(U)\}$ - initial size of network, $rg \in [0,1]$ - resolution of granulation,

Result: $IS' = (Y, A \cup \{a_{gr}\})$ - an information system, where the last attribute $a_{gr} : Y \to \{1,\ldots,nc\}$ denotes label of generated granule and $card(Y) \leq card(U)$ and $\forall x \in U \exists y \in Y \delta(x,y) < NR$

begin

```
[NR_init, Y] ⟵ initialize(U, A, size_net);
for y_i, y_j ∈ Y, i ≠ j do /*form clusters*/
    if δ(y_i, y_j) < NR_init then connect(y_i, y_j);
NR ⟵ NR_init;
while ¬stopIterations(Y) do
    for y ∈ Y do
        Δ(y) = (δ(y, x))_{x∈U}; /*calculate distances from input data*/;
        s_i(y) = NR − min Δ(y);/*similarity level of the object*/;
    delete(U, A, Y); /*remove redundant network objects*/;
    for y_i, y_j ∈ Y, i ≠ j do /* reconnect objects*/
        if δ(y_i, y_j) < NR then connect(y_i, y_j);
        a_gr(y_i) ⟵ 0; a_gr(y_j) ⟵ 0;
    grLabel ⟵ 1;
    for y_i ∈ Y do /*label objects*/
        if a_gr = 0 then a_gr(y_i) ⟵ grLabel;
        for y_j ∈ Y, j ≠ i do
            if connected(y_i, y_j) then
                a_gr(y_j) ⟵ grLabel;
        grLabel ⟵ grLabel + 1;
    for y_i ∈ Y do
        /*calculate the nearest neighbor network objects*/;
        δ_NN(y_i) = min({δ(y_i, y_j) : y_j ∈ Y & j ≠ i});
    NR ⟵ rg · Σ_{y∈Y} δ_NN(y) / card(Y) ;/*new value of NR*/;
    if ¬stopIterations(Y)/*test stopping condition */
    then
        joinNotRepresented(U, Y, NR, Δ);
        adjust(Y, U, A, NR);
```

Algorithm 2: Initial steps of SOSIG algorithm.

Data: Set of input objects: $U = \{x_1, \ldots, x_n\}$, $size_{net} \in \{0, 1, \ldots, card(U)\}$ - initial size of network
Result: Set of initial network objects $Y = \{y_1, \ldots, y_{size_{net}}\}$, where $Y \subset U$, NR_{init} - initial value of neighborhood radius threshold

```
begin
    δ_maxNN = 0;
    for x_i ∈ U do
        /*calculate the nearest neighbor
        distances of the data*/ ;
        δ_NN(x_i) = min({δ(x_i, x_j) : x_j ∈
        U & x_j ≠ x_i}) ;
        /*find the greatest value of the nearest
        neighbor distances*/ ;
        if δ_NN(x_i) > δ_maxNN then
        δ_maxNN = δ_NN(x_i);
    NR_init ⟵ δ_maxNN;
    /*select the representatives objects*/ ;
    for netObj ← 1 to size_net do
        i = rand(1, ..., card(U));
        y_netObj = x_i;
```

The measure of their usefulness is a similarity level expressed by Equation 5.

$$s_l(y) = NR - \min(\{\delta(y,x) : x \in U\}) \qquad (5)$$

The s_l defines a degree of similarity between examined representative point y and the most similar point x from the training data. There are considered only training points from neighborhood of y (defined by NR). The similarity level corresponds to the distance between the points - the closer points are more similar to each other than points located in farther distance. After calculation of similarity levels their values are normalized. In presented experiments to calculate distance between objects (hyperboxes) a distance measure expressed by Equation 6 was used:

$$d(I_A, I_B) = (\|a_B - a_A\| + \|b_B - b_A\|)/2 \qquad (6)$$

where $\|a_B - a_A\|$ and $\|a_B - a_A\|$ denote sum of subtractions of respectively minimal (a) and maximal (b) values of granules I_A and I_B in every dimension. The equation has been introduced in (Bargiela and Pedrycz, 2001).

To control size of the network there is a removal step as it is presented in Algorithm 3. Useless objects are removed. It affects redundant objects from the representatives network. As redundant are determined points with the same input object (from U) in their neighborhood. The best points stay in the network and also the ones that are not redundant for other input data. This process controls size of the network prevents forming excessively dense clusters. It results in compression phenomenon.

Algorithm 3: Detailed steps of SOSIG function *delete*.

Data: Set of input objects U, network set Y, set of attributes $A = \{a_1, \ldots, a_k\}$, NR - threshold of neighborhood radius
Result: $Y \setminus C$, where C is a set of redundant network objects

```
begin
    C ⟵ ∅;/*initially the set is empty*/;
    for x ∈ U do
        for y ∈ Y do
            if δ(y,x) < NR then /*add to the
            set objects representing the input
            element x*/;
            C ⟵ C ∪ {y};
        δ_NN(x, y_NN) = min({δ(x,y) : y ∈ Y}) ;
        C ⟵ C \ {y_NN};/*remove from the set
        the best object representing x */;
    Y ⟵ Y \ C;/*remove from the network
    objects from the set C*/;
```

The remaining objects in the network are reconnected and labeled. A granule is determined by edges between the objects in the structure. Components of the same granule (group) have equal label. Afterwards, a new value of NR parameter is calculated (see Equation 4). When stopping criterion is met, the algorithm is stopped after connections reconstruction and labeling. Otherwise, the following steps are carried out. As a stopping criterion, a stable state of the network is considered. That is the state of small fluctuations of network size and value of NR.

The last step is to apply a procedure of adjusting of all network objects (see Algorithm 5). In this step, candidate objects, that are copies of original ones, are created. The number of copies is not great, fixed in algorithm on 5. Values of attributes of candidate objects are slightly modified (depending on a similarity level of objects and the type of attributes). This procedure allows to adjust network objects in attained solution to the examined problem.

In the algorithm there is also a step of introducing an object from input data to the system. It concerns the object, which has not been identified yet. This operation (presented in Algorithm 4) avoids leaving

uncovered (not represented by network objects) area in the training set.

Further classification of new as well as training points can be performed using so-created structure (see Algorithm 6). To assign a label to the considered object it is necessary to determine neighborhood objects from the network structure. The neighborhood of the point is defined by final value of the NR (the last calculated value) of the SOSIG. The predominant value of labels is given to the examined object.

It must be underlined that the algorithm SOSIG does not require the number of clusters to be given. Partitioning is generated automatically, which eliminates inconvenient step of assessing and selecting the best result from a set of potential clusterings.

The algorithm SOSIG is also described in (Stepaniuk and Kużelewska, 2008).

Algorithm 4: Detailed steps of SOSIG function *joinNotRepresented*.

Data: Set of input objects: $U = \{x_1, \ldots, x_n\}$, $Y = \{y_1, \ldots, y_n\}$, Δ - matrix of distances between input and network objects, NR - threshold of neighborhood radius
Result: $Y \cup \{x\}$ (with condition $\neg\exists y \in Y \delta(y,x) < NR$)
begin
  for $x \in U$ **do**
    /*find an arbitrary object from the training set not represented yet by any network element*/ ;
    $add \longleftarrow 1$;
    for $y \in Y$ **do**
      if $\delta(y,x) < NR$ **then** $add \longleftarrow 0$;
      break;
    if $add = 1$ **then** $Y \longleftarrow Y \cup \{x\}$;
    break;

Algorithm 5: Detailed steps of SOSIG function *adjust*.

Data: Set of input objects U, network set Y, set of attributes $A = \{a_1, \ldots, a_k\}$, NR - threshold of neighborhood radius
Result: $Y \cup Z$, where Z is a set of adjusted network objects
begin
  $Z \longleftarrow \emptyset$;
  for $y \in Y$ **do**
    for $candidate \longleftarrow 1$ **to** $noCandidates$ **do**
      $z_{candidate} \longleftarrow y$;
      for $a \in A$ **do**
        $sign \longleftarrow rand(\{-1, 1\})$;
        $delta \longleftarrow sign$;
        if $a(z_{candidate})$ *is binary* **then**
          /*modification of binary values of attributes*/;
          $randVal \longleftarrow rand(\{0, 1\})$;
          $delta \longleftarrow delta \cdot randVal$;
          $a(z_{candidate}) \longleftarrow delta$;
        else
          /*modification of continuous values of attributes*/;
          $randVal \longleftarrow rand([0, 1])$;
          /*scale of change depends on the value of s_l */;
          $delta \longleftarrow delta \cdot randVal \cdot (1.5 - s_l(y))$;
          $a(z_{candidate}) \longleftarrow a(z_{candidate}) + delta$;
      for $x \in U$ **do**
        if $\delta(z_{candidate}, x) < NR$ **then**
          /*only useful clones are joined to the network */;
          $Z \longleftarrow Z \cup \{z_{candidate}\}$;
          break;

3 EXPERIMENTS

The article proposes a method for detection of groups containing similar objects. The method clusters data, that contains data points as well as hyperboxes. The experiments focus on comparing results of clustering in two approaches: when data are in point and granulated form. There were performed the following comparisons: the number of detected groups, values of SOSIG parameters, time of detection process, values of validity indices. It was also taken into consideration interpretability of created clusterings.

3.1 Description of the Datasets

There are several data sets in the experiments, shown in Table 1. The sets are various with regard to the number of objects, dimensionality and the existed number of groups. The column *groups numbers* from the Table 1 contains the number of groups presented in the data according to the subjective human perception based on the separation and compactness of the groups. However, the *irises* data set is a real data delivered with a priori class attribute. For that reason

Algorithm 6: Clustering of new objects in SOSIG algorithm.

Data:

- $IS = (U, A)$ - an information system, where $U = \{x_1, \ldots, x_n\}$ is a set of objects and $A = \{a_1, \ldots, a_k\}$ is a set of attributes,
- $IS' = (Y, A \cup \{a_{gr}\})$ where last attribute $a_{gr} : Y \rightarrow \{1, \ldots, ng\}$ stands for label of generated granule and $card(Y) \leq card(U)$
- NR - threshold of neighborhood radius;

Result: Clustered information system $IS_{gr} = (U, A \cup \{a_{gr}\})$ into ng clusters (granules), where last attribute $a_{gr} : U \rightarrow \{1, \ldots, nc\}$ stands for label of generated granule

begin

```
for x ∈ U do
    for granule ← 1 to ng do
        grLabels[granule] ← 0;
    /*calculate distance between x and the
    network objects*/;
    for y ∈ Y do
        if δ(x,y) < NR then
        label ← a_gr(y);
        grLabels[label] ←
        grLabels[label] + 1;
    /*predominant label is selected */;
    a_gr(x) ← max({grLabels}) ;
```

Table 1: Description of data sets.

data set	dim	objects number	hyperboxes number	groups number
norm2D2gr	2	200	51	2
sph2D6gr	2	300	70	6
irises	4	150	94	3
sph10D4gr	10	200	13	4

the value of the groups number for it is related to the number from the decision attribute.

3.2 Results of Experiments

Clustering by SOSIG algorithm was performed on both point-type and granulated data. The number of point-type and granular objects is compared in Table 1. It can be noticed that in all cases the number of hyperboxes is significantly less than the number of points. However the dimension is doubled for granulated data.

Number of groups identified by SOSIG in described above data sets is presented in Table 2. When the result consists of groups of highly various sizes, only a number of main groups is presented.

Table 2: Results of clustering of point-type and granulated data with respect to number of identified groups.

data set	point-type data	granulated data
norm2D2gr	2	2
sph2D6gr	6	6
irises	2, 3	4
sph10D4gr	4	4

Granulation of the *irises* set contains two levels (low and high resolution), what is visible in the all following tables. In the result are 2 clusters when granulation is performed on low resolution level, whereas in high resolution one large cluster is split in two smaller and there are additionally 5 significantly smaller groups. The results considering both levels of granulation are shown in the same cell of the tables, where first value corresponds to low and the second to high level of resolution. Granulation of *irises* hyperboxes is composed of only one level clustering with 4 main and 6 additional smaller groups.

In the remaining data sets clusterings the number of groups is corresponding to each other for both types of processed data.

Table 3: SOSIG results with respect to generated number of representatives and NR value when clustering point-type and granulated data.

data set	point data		granulated data		
	nr of reps.	NR val.	nr of repres.	NR val.	K val.
norm2D2gr	80	0.11	21	0.25	10
sph2D6gr	123	0.07	33	0.09	10
irises	92, 112	0.22, 0.1	88 88	0.13 0.13	15 15
sph10D4gr	46	0.21	6	0.18	6

The number of representatives generated by SOSIG and the NR value for point-type as well as hyperboxes data are shown in Table 3. Additionally, for granulated data K parameter is presented. K was selected to have small number of hyperboxes and preserving combining points from different clusters into one granule. For the *irises* point-type data set two levels of granulation were considered - composed of 2 and 3 groups. The number of representatives is less in case of clustering granulated data. Values of the NR are different in both cases, however there are comparable NR values for clustering of $sph2D6gr$, *irises* (version of high level clustering for point-type data) and $sph10D4gr$ sets.

Table 4: Average time (in seconds) of clustering granulated and point-type data.

data set	point-type data, t_{pd}	granulated data, t_{gd}	t_{pd}/t_{gd}
norm2D2gr	0.36	0.04	9
sph2D6gr	0.93	0.08	11.63
irises	0.87, 0.80	0.79	1.01
sph10D4gr	0.27	0.01	38.57

The results presented in Table 4 consider time (in seconds) of a single run of SOSIG algorithm. This is average time of 50 runs of the algorithm calculated for clustering original data as well as hyperboxes. The last column of the table contains quotient of the values. It can be seen the processing of granulated data is significantly, up to about 40 times, faster than processing original point-type objects. The most acceleration is visible when the number of objects in data is great.

To compare results of clustering one can use validity indices designed to detect the most compact and separable partitioning. The validity indices are not universal, however the most popular tool for assessment clustering results (Halkidi and Batistakis, 2001). Simultaneous comparison of several of them can give quite objective result.

Evaluation of groupings of granulated data in comparison to clusterings of point-type objects was performed in the experiments and the results are shown in Table 5. The following indices were taken into consideration: Davies-Bouldin's (DB), Dunn's, Silhouette (SI) and CDbw. Clusterings of *norm2D2gr*, *sph2D6gr* and *sph10D4gr* in form of hyperboxes are characterized by better values of 3 from 4 indices. For the set *irises* the values are better for point-type clustering.

Table 6 contains detailed description of groups (granules) detected in clustering of *irises* hyperboxes data. The final result is composed of 10 clusters, however due to considerable differences in their size, the result focuses on the main 3 granules. A priori decision attribute is composed of 3 classes: Iris-setosa (I-S), Iris-versicolor (I-Ve) and Iris-virginica (I-Vi). The set is described by 4 attributes: sepal-length (SL), sepal-width (SW), petal-length (PL) and petal-width (PW). The granule gr_1 contains 13 smaller granules (hyperboxes) and all of them belong to the class Iris-setosa. The other granule (gr_3) has comparable size (15 objects) and contains only objects from the Iris-versicolor class. The largest granule, gr_2 consists of 36 hyperboxes. It is not homogenous with respect of class attribute, due to 31% of objects come from Iris-versicolor class and 69% from Iris-virginica.

Table 5: Results of clustering (values of the indices) of point-type and granulated data.

data set	indices	point-type data	granulated data
norm2D2gr	DB	0.06	0.08
	Dunn's	0.16	0.74
	SI	0.53	0.76
	CDbw	19.15	45.34
sph2D6gr	DB	0.03	0.01
	Dunn's	0.51	1.38
	SI	0.75	0.85
	CDbw	36.54	23.58
irises	DB	0.14, 0.12	0.2
	Dunn's	0.39, 0.19	0.25
	SI	0.63, 0.53	0.3
	CDbw	74.84, 2.91	1.61
sph10D4gr	DB	0.01	0.0001
	Dunn's	7.83	9.29
	SI	0.93	0.96
	CDbw	453.23	41.04
	DB	0.23	0.37

It has to be focused attention on the attributes resulted from doubling of dimensions. These features are related to minimal and maximal values of the original attributes. As a consequence it appears an additional feature - difference between the maximal and minimal value of particular variables. Average differences are presented in Table 6 in column $diff_{Avg}$. The granule gr_1 is characterized by the widest range of all attributes, the granule gr_2 contains flowers with the smallest size of petals and sepals. Finally, the granule gr_3 is composed of irises with narrow and long petals and sepals.

Table 7 presents granules from the second level of data relationship hierarchy. The granules are hyperboxes identified in the first phase of the granulation. In the table the greatest 3 hyperboxes (denoted as gr_{ij}) from every granule of the main level were selected. The second-level granules from the top-level granules gr_1 and gr_3 have larger size and the range of their attributes values is greater on contrary to the granules belonging to gr_2. It shows the granules gr_1 and gr_3 are more compact and have greater regions of even information density. It can be noticed, that the hyperboxes are homogenous with regard to class attribute.

4 CONCLUSIONS

The article presents modified clustering method as an approach for data granulation. The algorithm is two-

Table 6: Main level of *irises* data hierarchy composed of clustering result of hyperboxes set. Table contains 3 main granules.

id/ size	class distr.	attr.	min val.	max val.	$diff_{Avg}$
gr_1/ 13	100% I-S	SL	4.4-5.4	4.8-5.5	0.25
		SW	3.0-3.7	3.1-3.9	0.15
		PL	1.0-1.5	1.5-1.9	0.29
		PW	0.1-0.4	0.1-0.5	0.12
gr_2/ 36	31% I-Ve 69% I-Vi	SL	5.6-7.1	5.6-7.1	0.04
		SW	2.5-3.4	2.5-3.4	0.03
		PL	4.3-6.0	4.4-6.0	0.07
		PW	1.4-2.5	1.4-2.5	0.02
gr_3/ 15	100% I-Ve	SL	5.2-6.1	5.2-6.2	0.11
		SL	5.2-6.1	5.2-6.2	0.11
		SW	2.3-2.9	2.3-3.0	0.08
		PL	3.5-4.7	3.6-4.7	0.17
		PW	1.0-1.4	1.1-0.5	0.08

Table 7: Second level of *irises* data hierarchy composed of hyperboxes (there are presented only selected objects).

main granule	gr. id	size	class distr.	min values of attr.		$diff_{Avg}$
gr_1	gr_{11}	15	100% I-S	SL	5.0	0.5
				SW	3.4	0.3
				PL	1.3	0.4
				PW	0.2	0.2
	gr_{12}	15	100% I-S	SL	4.6	0.5
				SW	3.3	0.3
				PL	1.0	0.7
				PW	0.2	0.3
	gr_{13}	9	100% I-S	SL	4.8	0.2
				SW	3.0	0.2
				PL	1.2	0.4
				PW	0.1	0.2
gr_2	gr_{21}	5	100% I-Ve	SL	6.4	0.3
				SW	2.9	0.2
				PL	4.3	0.4
				PW	1.3	0.2
	gr_{22}	4	100% I-Vi	SL	6.4	0.1
				SW	3.0	0.2
				PL	5.1	0.4
				PW	1.8	0.2
	gr_{23}	4	100% I-Vi	SL	5.9	0.3
				SW	2.8	0.2
				PL	4.8	0.3
				PW	1.8	0.0
gr_3	gr_{31}	14	100% I-Ve	SL	5.6	0.5
				SW	2.7	0.3
				PL	3.9	0.8
				PW	1.2	0.3
	gr_{32}	8	100% I-Ve	SL	5.7	0.5
				SW	2.6	0.3
				PL	3.5	0.8
				PW	1.0	0.3
	gr_{33}	6	100% I-Vi	SL	5.4	0.3
				SW	2.8	0.2
				PL	4.1	0.4
				PW	1.3	0.2

phased, first phase prepares input point-type data as multi-dimensional granules in form of hyperboxes. The hyperboxes are based on maximizing information density in data. The next phase is clustering the granules by SOSIG algorithm. Clustering process can be performed on different resolution of data. Clustering of hyperboxes was executed without changing of resolution. Three-level structure of data was constructed by joining original point (third down level) in hyperboxes (second level), whereas the top level contains division of hyperboxes into clusters. Partitioning at the top level of hyperboxes granulation (clustering) is composed of the same number of groups as partitioning point-type data. Quality of created clusters is comparable as well, due to values of quality indices are similar.

Process of hyperboxes creation is a type of aggregation operation, therefore the most benefit of the presented method is shortening time of clusters creation in comparison to the processing point-type data. It is especially effective when data contain large number of objects. Hyperboxes also determine additional level of relationship existing in data. Finally, description of the granules is more comprehensible since the hyperboxes contain minimal and maximal values of attributes.

ACKNOWLEDGEMENTS

This work was supported by Rector's of Technical University of Bialystok Grant No. S/WI/5/08.

The experiments were performed on the computer cluster at Faculty of Computer Science, Bialystok Technical University.

REFERENCES

A.K. Jain, M. M. and Flynn, P. (1999). Data clustering: a review. In *ACM Computing Surveys 31:3, 264–323*.

Bargiela, A. and Pedrycz, W. (2001). Classification and clustering of granular data. In *IFSA World Congress and 20th NAFIPS International Conference, vol.3, 1696–1701*.

Bargiela, A. and Pedrycz, W. (2002). *Granular Computing: an Introduction*. Kluwer Academic Publishers, Boston.

Bargiela, A. and Pedrycz, W. (2006). Granular analysis of traffic data for turning movements estimation. In *Int. J. of Enterprise Information Systems, vol. 2-2, 13–27*.

Halkidi, M. and Batistakis, Y. (2001). On clustering validation techniques. In *Journal of Intelligent Information Systems 17:2/3, 107–145*.

Halkidi, M. and Vazirgiannis, M. (2002). Clustering validity assessment using multi representatives. In *Proceedings of SETN Conference*.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.

Pawlak, Z. (1991). *Rough Sets. Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht.

Stepaniuk, J. and Kużelewska, U. (2008). Information granulation: A medical case study. In *Transactions on Rough Sets, vol.5390/2008, 96–113*. Springer.

Wierzchoń, S. and Kużelewska, U. (2006). Evaluation of clusters quality in artificial immune clustering system - saris. In *Biometrics, computer security systems and artificial intelligence applications,323–331*. Springer-Verlag.

Yao, Y. (2006). Granular computing for data mining. In *Proceedings of SPIE Conference on Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security,1–12*.

Zadeh, L. A. (2001). A new direction in ai: Toward a computational theory of perceptions. In *AI Magazine 22(1), 73-84*.