Int. J. Appl. Math. Comput. Sci., 2020, Vol. 30, No. 3, 561–572
DOI: 10.34768/amcs-2020-0041

ANT–BASED CLUSTERING FOR FLOW GRAPH MINING

ARKADIUSZ LEWICKI [a,*], KRZYSZTOF PANCERZ [b]

[a] Chair of Applied Information Systems
University of Information Technology and Management in Rzeszów
Sucharskiego 2, 35-225 Rzeszów, Poland
e-mail: alewicki@wsiz.rzeszow.pl

[b] Institute of Computer Science
University of Rzeszów
Pigonia 1, 35-310 Rzeszów, Poland
e-mail: kpancerz@ur.edu.pl

The paper is devoted to the problem of mining graph data. The goal of this process is to discover possibly certain sequences appearing in data. Both rough set flow graphs and fuzzy flow graphs are used to represent sequences of items originally arranged in tables representing information systems. Information systems are considered in the Pawlak sense, as knowledge representation systems. In the paper, an approach involving ant based clustering is proposed. We show that ant based clustering can be used not only for building possible large groups of similar objects, but also to build larger structures (in our case, sequences) of objects to obtain or preserve the desired properties.

Keywords: possibly certain sequences, flow graphs, rough sets, fuzzy sets, ant-based clustering.

1. Introduction

Modeling behavior of systems, including biological, technological, medical, economical, sociological, and psychological ones, by means of diverse soft computing tools is one of the popular tasks in computer science research (Marwala, 2013; Tadeusiewicz, 2015). On the basis of such models, a sequence (including temporal) data mining becomes an important issue (Dong and Pei, 2007; Kumar *et al.*, 2011; Mitsa, 2010). There is huge literature concerning this topic. In many domains, sequences of objects (states, events or other representations of phenomena) are collected. Discovering valuable knowledge from such sets of sequences is an important task in many applications, e.g., behavior analysis, gene analysis, process mining.

In the presented approach, we are interested in sequences that are ordered objects, particularly objects recorded in consecutive time instants. Mining data to discover interesting sequences is one of the common tasks among machine learning problems. On the one hand, most studies consider mining simple or complex sequences to find the so-called frequent episodes (Huang and Chang, 2008; Mannila *et al.*, 1997), i.e., collections of events occurring frequently together. On the other hand, unique sequences, i.e., sequences which clash, in some sense (for example, according to a given criterion), with other sequences in the set, are discovered (Pancerz *et al.*, 2012).

In the presented approach, we are interested in possibly certain sequences appearing in data. Possibly certain sequences are sequences with certainties as high as possible. We assume that data are arranged in the form of data tables representing information systems in the Pawlak sense, i.e., a knowledge representation systems (Pawlak, 1991). The certainties of sequences are determined on the basis of flow graphs (either rough set flow graphs or fuzzy flow graphs) build for information systems. In general, an information flow distribution is a kind of knowledge that can be helpful in solving different problems appearing in data analysis, especially, if we deal with ordered data, i.e., data constituting sequences of objects. In the literature, different approaches based on flow graphs were proposed. The fundamental one, called

*Corresponding author

flow networks, was proposed by Ford and Fulkerson (2010). In the paper, we are interested in two other approaches introduced in the area of data mining, namely, fuzzy flow graphs proposed by Mieszkowicz-Rolka and Rolka (2006) and flow graphs (called here rough set flow graphs) proposed by Pawlak (2005).

Our approach differs from the approaches to discovering frequent episodes. Firstly, episodes are considered as collections of events that occur relatively close to each other in a given partial order (Mannila *et al*., 1997). Among others, serial and parallel episodes are considered. In our approach, we are interested in ordered (not partially) sets of attribute values or linguistic values describing objects (cases) of interest. Secondly, our approach is focused on certainties of sequences, not on their frequencies of occurrence. It is worth noting that the idea implemented in our approach differs from sequence clustering which can be called vertical clustering (sequences are clustered to build groups of sequences). Meanwhile, our clustering can be called horizontal clustering. Shorter pieces of sequences (one-element sequences at the beginning) are joined to build longer ones (in order to get as high certainty as possible).

In real life data, the space for all possible sequences is very large. Therefore, we propose to use the ant-based clustering procedure that represents one of the heuristic approaches in discovering possibly certain sequences. Ant-based clustering is a biologically inspired data clustering technique (see Deneubourg *et al*., 1991; Handl *et al*., 2006; Lumer and Faieta, 1994). An important contribution of our paper is to show that ant based clustering can be used not only for building possible large groups of similar objects but also to build larger structures (in our case, sequences) of objects to obtain or preserve the desired properties (Parpinelli *et al*., 2002; Pancerz *et al*., 2015).

The remaining part of the paper is organized as follows. Section 2 provides a thorough description of the theoretical background for each aspect of the proposed approach. Definitions of significant notions are explained by simple examples. In Section 3, the ant-based clustering procedure for discovering possibly certain sequences (created over the sets of attribute values or created over the sets of linguistic values associated with linguistic variables defined for attributes in information systems) is presented. Moreover, the description of results of experiments on real life data is included in Section 3. Finally, Section 4 consists of some conclusions and directions for further work.

2. Theoretical background

2.1. Information systems. In the approach presented in this paper, information systems are understood as Pawlak's knowledge representation systems (Pawlak, 1991). In this sense, information systems are a mathematical tool for describing some objects, cases, phenomena we are intended to analyze, classify, group, etc.

Definition 1. An *information system* IS is the quadruple

$$\text{IS} = (U, A, \{V_a\}_{a\in A}, f_{\text{inf}}),$$

where

- U is a nonempty, finite set of objects,
- A is a nonempty, finite set of attributes,
- $\{V_a\}_{a\in A}$ is a family of nonempty sets of attribute values,
- $f_{\text{inf}} : A \times U \to \bigcup_{a\in A} V_a$ is an information function such that $f_{\text{inf}}(a, u) \in V_a$ for each $a \in A$ and $u \in U$.

Any information system can be presented as a data table. The columns of the table are labeled with attributes from the set A, the rows are labeled with objects from the set U, and the entries of the table are the values of the information function f_{inf} assigning to each attribute $a \in A$ and each object $u \in U$ a value of a on u.

We can consider an information system IS $= (U, A, \{V_a\}_{a\in A}, f_{\text{inf}})$ in which a set A of attributes is ordered (particularly ordered in time). In this case, a set A of attributes in IS is, in fact, a sequence of attributes, i.e., $A = \langle a_t : t = 1, 2, \ldots, m\rangle$, where a_t is the attribute determining values of objects from U at (time)point t. Further, we will be interested in such information systems.

In the case of information systems with the ordered sets of attributes, we can consider sequences of attribute values.

Definition 2. Let IS $= (U, A, \{V_a\}_{a\in A}, f_{\text{inf}})$ be an information system, where $A = \langle a_t : t = 1, 2, \ldots, m\rangle$. $s(j) = \langle v_{s_1}, v_{s_2}, \ldots, v_{s_k}\rangle$, $k \leq m$, is a sequence over the sets of attribute values in IS, i.e., $v_{s_1} \in V_{a_j}$, $v_{s_2} \in V_{a_{j+1}}$, $\ldots$, $v_{s_k} \in V_{a_{j+k-1}}$, where $j \in \{1, 2, \ldots, m-k\}$, starting at point j. The *cardinality* card$(s(j))$ of a sequence $s(j)$ will be denoted by $|s(j)|$.

Remark 1. Let $s(j)$ be a sequence over the sets of attribute values in a given information system IS starting at any point j. If $|s(j)| = 1$, then the sequence $s(j)$ is called a degenerated sequence.

Example 1. Consider a simple information system $\text{IS}^1 = (U^1, A^1, \{V_a\}_{a\in A^1}, f^1_{\text{inf}})$, where $U^1 = \{u_1, u_2, \ldots, u_{10}\}$, $A^1 = \langle a_1, a_2, a_3\rangle$, $V_{a_1} = \{X, Y, Z\}$, $V_{a_2} = \{W, Y, Z\}$, and $V_{a_3} = \{X, Y\}$. This information system is presented as a data table in Table 1. The sequences

- $s_1(1) = \langle Z, Z, Y\rangle$,
- $s_2(1) = \langle X, Y\rangle$,

Table 1. Information system IS^1 presented as a data table.

U^1 A^1	a_1	a_2	a_3
u_1	X	W	X
u_2	Y	W	X
u_3	X	Y	X
u_4	Z	Y	Y
u_5	X	Y	X
u_6	X	Y	Y
u_7	X	Y	Y
u_8	Y	Z	X
u_9	Y	Z	Y
u_{10}	Z	Z	Y

are sequences over the sets of attribute values in IS^1. One can see that $|s_1(1)| = 3$ and $|s_2(1)| = 2$. ♦

There are two key types of attribute values in information systems: numerical and symbolic. Numerical values are expressed by numbers (e.g., real numbers, integers, prime numbers, etc.). Symbolic values usually describe qualitative concepts. Let $\mathbb{R}$ be a set of real numbers. A numerical attribute in an information system is an attribute whose set of values is a non-empty subset of $\mathbb{R}$. A symbolic attribute in an information system is an attribute whose set of values includes symbolic values only.

2.2. Fuzzification. Fuzzification is the process that transforms the real value variables into linguistic variables whose domains contain linguistic values which can be described by fuzzy sets (their membership functions).

Now, we are interested in information systems with numerical attributes only. Let $IS = (U, A, \{V_a\}_{a\in A}, f_{\text{inf}})$ be an information system such that $V_a \subseteq \mathbb{R}$ for each $a \in A$. For each attribute $a \in A$, we can define a linguistic variable λ_a. With each linguistic variable λ_a, a set L^{λ_a} of linguistic values is associated:

$$L^{\lambda_a} = \{l_1^a, l_2^a, \ldots, l_{k_a}^a\}.$$

Each linguistic value l_i^a, where $i = 1, 2, \ldots, k_a$, is described by a membership function $\mu_{l_i^a} : \mathbb{R} \to [0, 1]$. In the literature, a lot of different membership functions have been defined to describe linguistic values (e.g., triangular, trapezoidal).

2.3. Fuzzified information systems. For an information system with numerical attributes only, we can create a fuzzified information system as a result of the application of fuzzification processes for sets of attribute values.

Definition 3. Let

- $IS = (U, A, \{V_a\}_{a\in A}, f_{\text{inf}})$ be an information system with $U = \{u_1, u_2, \ldots, u_n\}$ and $A = \{a_1, a_2, \ldots, a_m\}$, such that $V_a \subseteq \mathbb{R}$ for each $a \in A$,
- $\{L^{\lambda_a}\}_{a\in A}$ be the family of sets of linguistic values associated with linguistic variables from the family $\{\lambda_a\}_{a\in A}$ defined for attributes from A, where $L^{\lambda_a} = \{l_1^a, l_2^a, \ldots, l_{k_a}^a\}$ for each $a \in A$.

A *fuzzified information system* $\mathcal{F}(\text{IS})$ corresponding to IS, is the quadruple

$$\mathcal{F}(\text{IS}) = (U^{\mathcal{F}}, \Phi, \{V_\phi\}_{\phi\in\Phi}, f_{\text{inf}}^{\mathcal{F}}),$$

where

- $U^{\mathcal{F}}$ is a nonempty, finite set of objects such that each $u^* \in U^{\mathcal{F}}$ corresponds exactly to one $u \in U$,
- $\Phi = \Phi_{a_1} \cup \Phi_{a_2} \cup \cdots \cup \Phi_{a_m}$ is the nonempty, finite set of fuzzified attributes, such that
  - $\Phi_{a_1} = \{a_1^{l_1^{a_1}}, a_1^{l_2^{a_1}}, \ldots, a_1^{l_{k_{a_1}}^{a_1}}\}$,
  - $\Phi_{a_2} = \{a_2^{l_1^{a_2}}, a_2^{l_2^{a_2}}, \ldots, a_2^{l_{k_{a_2}}^{a_2}}\}$,
  - $\ldots$,
  - $\Phi_{a_m} = \{a_m^{l_1^{a_m}}, a_m^{l_2^{a_m}}, \ldots, a_m^{l_{k_{a_m}}^{a_m}}\}$,
- $\{V_\phi\}_{\phi\in\Phi}$ is a family of sets of fuzzified attribute values and $V_\phi = [0, 1]$ for each $\phi \in \Phi$,
- $f_{inf}^{\mathcal{F}} : \Phi \times U^{\mathcal{F}} \to \bigcup_{\phi\in\Phi} V_\phi$ is the information function such that
  - $f_{\text{inf}}^{\mathcal{F}}(a^{l_i^a}, u^*) \in V_\phi$ for each $a^{l_i^a} \in \Phi$ and $u^* \in U^{\mathcal{F}}$,
  - $f_{\text{inf}}^{\mathcal{F}}(a^{l_i^a}, u^*) = \mu_{l_i^a}(f_{\text{inf}}(a, u))$, where $\mu_{l_i^a}$ is a membership function describing l_i^a and $u^* \in U^{\mathcal{F}}$ corresponds to $u \in U$,

  for each $a \in A$ and $i = 1, 2, \ldots, k_a$.

If some attributes of an information system are symbolic (this situation is common for decision attributes), then we can use the so-called binary fuzzification for them.

If, for a given $a \in A$, the value set of a is a finite set $V_a = \{v_1, v_2, \ldots, v_{k_a}\}$ of symbolic values, then

- $\Phi_a = \{a^{v_1}, a^{v_2}, \ldots, a^{v_{k_a}}\}$,
- $f_{\text{inf}}^{\mathcal{F}}(a^{v_i}, u^*) = \begin{cases} 1, & f_{\text{inf}}(a, u) = v_i, \\ 0, & f_{\text{inf}}(a, u) \neq v_i, \end{cases}$

where $u^* \in U^{\mathcal{F}}$ corresponds to $u \in U$ and $i = 1, 2, \ldots, k_a$. One can see that in this case we use, in fact, the so-called fuzzy singleton membership function.

In the case of information systems being fuzzified, in which the sets of attributes are ordered, we can define sequences of linguistic values.

Table 2. Information system IS^2 presented as a data table.

U^2 A^2	a_1	a_2	a_3
u_1	0.2	3.1	4.7
u_2	0.5	1.0	2.8
u_3	1.8	4.4	2.0
u_4	3.4	4.3	0.9
u_5	3.5	2.5	0.4
u_6	0.6	1.2	2.6
u_7	0.7	0.9	3.4
u_8	2.4	2.9	4.9
u_9	0.8	0.8	2.0
u_{10}	3.5	2.1	0.3

Definition 4. Let $\text{IS} = (U, A, \{V_a\}_{a\in A}, f_{\text{inf}})$ be an information system, where $A = \langle a_t : t = 1, 2, \ldots, m\rangle$, and let $\{L^{\lambda_a}\}_{a\in A}$ be a family of sets of linguistic values associated with linguistic variables from the family $\{\lambda_a\}_{a\in A}$ defined for attributes from A, where $L^{\lambda_a} = \{l_1^a, l_2^a, \ldots, l_{k_a}^a\}$ for each $a \in A$. $s(j) = \langle v_{s_1}, v_{s_2}, \ldots, v_{s_k}\rangle$, $1 < k \leq m$, is a sequence over the sets of linguistic values associated with linguistic variables defined for attributes from A, i.e., $v_{s_1} \in L^{\lambda_{a_j}}$, $v_{s_2} \in L^{\lambda_{a_{j+1}}}$, $\ldots$, $v_{s_k} \in L^{\lambda_{a_{j+k-1}}}$, where $j \in \{1, 2, \ldots, m-k\}$, starting at point j.

It is worth noting that, referring to Remark 1, we can also consider degenerated sequences over the sets of linguistic values associated with linguistic variables defined for attributes in a given information system.

Example 2. Consider a simple information system $\text{IS}^2 = (U^2, A^2, \{V_a\}_{a\in A^2}, f_{\text{inf}}^2)$, where $U^1 = \{u_1, u_2, \ldots, u_{10}\}$, $A^2 = \langle a_1, a_2, a_3\rangle$. This information system is presented as a data table in Table 2.

For each attribute $a \in A^2$, we have defined a linguistic variable λ_a with a set L^{λ_a} of linguistic values $L^{\lambda_a} = \{\text{low}, \text{medium}, \text{high}\}$. An example of the fuzzified information system $\mathcal{F}(\text{IS}^2) = (U^{\mathcal{F}}, \Phi, \{V_\phi\}_{\phi\in\Phi}, f_{\text{inf}}^{\mathcal{F}})$ corresponding to the information system IS^2 is presented as a data table in Table 3.

The following sequences:

- $s_1(1) = \langle\text{low, low, medium}\rangle$,
- $s_2(2) = \langle\text{low, high}\rangle$

are sequences over the sets of linguistic values associated with linguistic variables defined for attributes from A^2. ♦

2.4. Triangular norms. For sets of real numbers within the interval $[0, 1]$, a special class of functions, called triangular norms, is considered. Triangular norms can be either t-norms or t-conorms (Klement *et al.*, 2000).

Definition 5. A *t-norm* is a function $T : [0, 1] \times [0, 1] \to [0, 1]$ such that, for $x, y, w, z \in [0, 1]$ the following conditions are satisfied:

1. $T(x, 1) = x$, $T(x, 0) = 0$,
2. $T(x, y) = T(y, x)$,
3. $T(x, T(y, z)) = T(T(x, y), z)$,
4. if $x \leq y$ and $w \leq z$, then $T(x, w) \leq T(y, z)$.

Definition 6. A *t-conorm* is a function $S : [0, 1] \times [0, 1] \to [0, 1]$ such that, for $x, y, w, z \in [0, 1]$ the following conditions are satisfied:

1. $S(x, 1) = 1$, $S(x, 0) = x$,
2. $S(x, y) = S(y, x)$,
3. $S(x, S(y, z)) = S(S(x, y), z)$,
4. if $x \leq y$ and $w \leq z$, then $S(x, w) \leq S(y, z)$.

In the literature, numerous different triangular norms have been defined. The most popular ones are as follows:

- t-norms:
  - Zadeh's t-norm,
$$T_Z(x, y) = \min(x, y),$$
  - algebraic t-norm,
$$T_A(x, y) = xy,$$
  - Lukasiewicz's t-norm,
$$T_L(x, y) = \max(x + y - 1, 0),$$
  - Einstein's t-norm,
$$T_E(x, y) = \frac{xy}{2 - (x + y - xy)},$$
- t-conorms:
  - Zadeh's t-conorm,
$$S_Z(x, y) = \max(x, y),$$
  - probabilistic t-conorm,
$$S_P(x, y) = x + y - xy,$$
  - Lukasiewicz's t-conorm,
$$S_L(x, y) = \min(x + y, 1),$$
  - Einstein's t-conorm,
$$S_E(x, y) = \frac{x + y}{1 + xy},$$

where $x, y \in [0, 1]$.

A triangular norm (either t-norm or t-conorm), will be generally denoted by $F(x, y)$, where $x, y \in [0, 1]$. Moreover, we will use the notation $F(x_1, x_2, \ldots, x_r)$ for $F(x_1, F(x_2, \ldots, F(x_{r-1}, x_r))$, where $x_1, x_2, \ldots, x_r \in [0, 1]$.

Table 3. Fuzzified information system $\mathcal{F}(\text{IS}^2)$ corresponding to the information system IS^2 presented as a data table.

$U^{\mathcal{F}}$ Φ	a_1^{low}	a_1^{medium}	a_1^{high}	a_2^{low}	a_2^{medium}	a_2^{high}	a_3^{low}	a_3^{medium}	a_3^{high}
u_1^*	1.0000	0.0000	0.0000	0.0000	0.6000	0.5500	0.0000	0.0000	1.0000
u_2^*	1.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.1000	0.8000	0.4000
u_3^*	0.6000	0.5333	0.0000	0.0000	0.0000	1.0000	0.5000	0.6667	0.0000
u_4^*	0.0000	0.4000	0.7000	0.0000	0.0000	1.0000	1.0000	0.0000	0.0000
u_5^*	0.0000	0.3333	0.7500	0.2500	1.0000	0.2500	1.0000	0.0000	0.0000
u_6^*	1.0000	0.0000	0.0000	0.9000	0.1333	0.0000	0.2000	0.9333	0.3000
u_7^*	1.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.4000	0.7000
u_8^*	0.3000	0.9333	0.2000	0.0500	0.7333	0.4500	0.0000	0.0000	1.0000
u_9^*	1.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.5000	0.6667	0.0000
u_{10}^*	0.0000	0.3333	0.7500	0.4500	0.7333	0.0500	1.0000	0.0000	0.0000

2.5. Rough set flow graphs. Rough set flow graphs were defined by Pawlak (2005) as a tool for reasoning from data.

Definition 7. Let $\text{IS} = (U, A, \{V_a\}_{a\in A}, f_{\text{inf}})$ be an information system with $U = \{u_1, u_2, \ldots, u_n\}$ and $A = \{a_1, a_2, \ldots, a_m\}$, such that $V_a = \{v_a^1, v_a^2, \ldots, v_a^{k_a}\}$ for each $a \in A$. A *rough set flow graph* corresponding to IS is the quintuple

$$\mathcal{RSFG}(IS) = (N, B, \text{cer}, \text{str}, \text{cov}),$$

where

- $N = N_{a_1} \cup N_{a_2} \cup \cdots \cup N_{a_m}$ is a set of nodes such that for each $a \in \{a_1, a_2, \ldots, a_m\}$: $N_a = \{\widehat{a}^{v_a^1}, \widehat{a}^{v_a^2}, \ldots, \widehat{a}^{v_a^{k_a}}\}$,
- $B \subseteq N \times N$ is a set of multi-labeled directed branches such that for any $(n^x, n^y) \in B$, $n^x \in N_{a_{i-1}}$ and $n^y \in N_{a_i}$ and $i \in \{2, 3, \ldots, m\}$,
- cer : $B \to [0, 1]$ is a certainty function labeling branches such that

$$\text{cer}(\widehat{a}_{i-1}^{v_{a-1}^x}, \widehat{a}_i^{v_a^y}) = \frac{\text{card}(\{u\in U: f_{\text{inf}}(a_{i-1},u)=v_{a_{i-1}}^x \wedge f_{\text{inf}}(a_i,u)=v_{a_i}^y\})}{\text{card}(\{u\in U: a_{i-1}(u)=v_{a_{i-1}}^x\})},$$

for any $(\widehat{a}_{i-1}^{v_{a-1}^x}, \widehat{a}_i^{v_a^y}) \in B$,

- str : $B \to [0, 1]$ is a strength function labeling branches such that

$$\text{str}(\widehat{a}_{i-1}^{v_{a-1}^x}, \widehat{a}_i^{v_a^y}) = \frac{\text{card}(\{u\in U: f_{\text{inf}}(a_{i-1},u)=v_{a_{i-1}}^x \wedge f_{\text{inf}}(a_i,u)=v_{a_i}^y\})}{\text{card}(U)},$$

for any $(\widehat{a}_{i-1}^{v_{a-1}^x}, \widehat{a}_i^{v_a^y}) \in B$,

- cov : $B \to [0, 1]$ is a covering function labeling branches such that

$$\text{cov}(\widehat{a}_{i-1}^{v_{a-1}^x}, \widehat{a}_i^{v_a^y}) = \frac{\text{card}(\{u\in U: f_{\text{inf}}(a_{i-1},u)=v_{a_{i-1}}^x \wedge f_{\text{inf}}(a_i,u)=v_{a_i}^y\})}{\text{card}(\{u\in U: a_i(u)=v_{a_i}^y\})},$$

for any $(\widehat{a}_{i-1}^{v_{a-1}^x}, \widehat{a}_i^{v_a^y}) \in B$.

One can see that we can distinguish particular layers in the set N of nodes of $\mathcal{RSFG}(\text{IS})$. The layer N_a, where $a \in \{a_1, a_2, \ldots, a_m\}$, corresponds exactly to one attribute $a \in A$. Each node in the layer N_a corresponds exactly to one value from the set V_a of values of a.

Example 3. Let us return to an information system IS^1 considered in Example 1. The rough set flow graph corresponding to the information system IS^1 visualized using the Graphviz tool (Ellson *et al.*, 2004) is shown in Fig. 1. ♦

On the basis of a rough set flow graph $\mathcal{RSFG}(\text{IS})$ corresponding to a given information system IS, we can calculate certainties of non-degenerated sequences over the sets of attribute values in IS.

Definition 8. Let

- $\mathcal{RSFG}(\text{IS}) = (N, B, \text{cer}, \text{str}, \text{cov})$ be a rough set flow graph corresponding to an information system $\text{IS} = (U, A, \{V_a\}_{a\in A}, f_{\text{inf}})$, where $A = \langle a_t : t = 1, 2, \ldots, m\rangle$,
- $s(j) = \langle v_{s_1}, v_{s_2}, \ldots, v_{s_k}\rangle$, be a sequence starting at point j, where $j < m$, over the sets of attribute values in IS, $|s(j)| > 1$.

The *certainty* $\text{cer}(s(j))$ of the sequence $s(j)$ is defined as

$$\text{cer}(s(j)) = F(\text{cer}(\widehat{a}_j^{v_{s_1}}, \widehat{a}_{j+1}^{v_{s_2}}), \text{cer}(\widehat{a}_{j+1}^{v_{s_2}}, \widehat{a}_{j+2}^{v_{s_3}}), \ldots, \text{cer}(\widehat{a}_{j+k-2}^{v_{s_{k-1}}}, \widehat{a}_{j+k-1}^{v_{s_k}})),$$

where F is a given triangular norm (see Section 2.4).

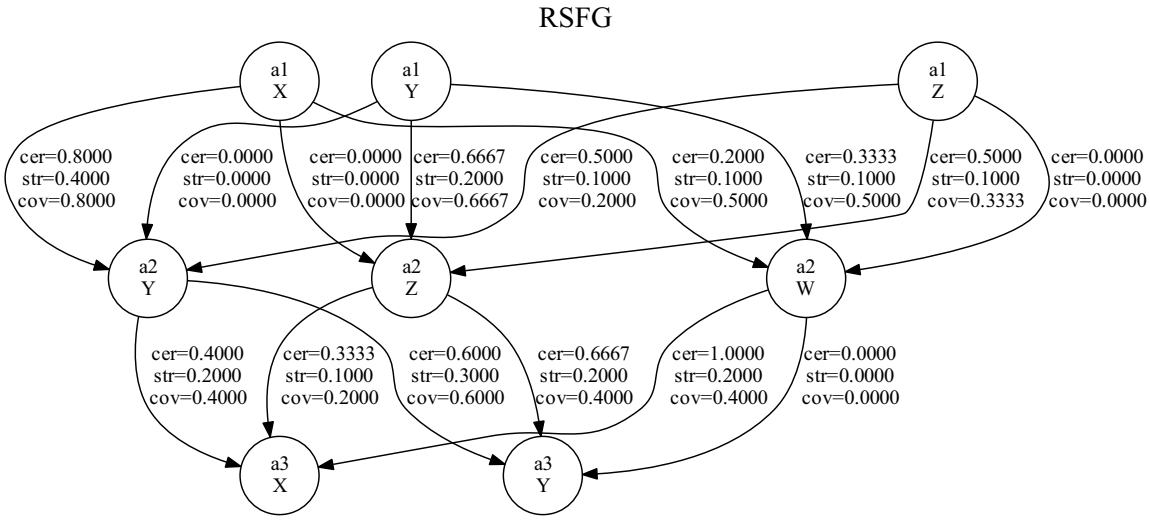

Fig. 1. Rough set flow graph corresponding to the information system IS^1.

Example 4. For the sequences considered in Example 1, one can see that

- $\text{cer}(\widehat{a}_1^Z, \widehat{a}_2^Z) = 0.5$,
- $\text{cer}(\widehat{a}_2^Z, \widehat{a}_3^Y) = 0.6667$,
- $\text{cer}(\widehat{a}_1^X, \widehat{a}_2^Y) = 0.8$.

Hence, we obtain the certainties of the considered sequences as collected in Table 4. ♦

2.6. Fuzzy flow graphs. Fuzzy flow graphs were proposed by Mieszkowicz-Rolka and Rolka (2006) to allow representation of information/decision tables with fuzzy attributes.

Definition 9. Let $\mathcal{F}(\text{IS}) = (U^{\mathcal{F}}, \Phi, \{V_\phi\}_{\phi\in\Phi}, f_{\text{inf}}^{\mathcal{F}})$ be a fuzzified information system corresponding to an information system $\text{IS} = (U, A, \{V_a\}_{a\in A}, f_{\text{inf}})$ with $U = \{u_1, u_2, \ldots, u_n\}$ and $A = \{a_1, a_2, \ldots, a_m\}$. A *fuzzy flow graph* corresponding to $\mathcal{F}(\text{IS})$ is the triple

$$\mathcal{FFG}(\mathcal{F}(\text{IS})) = (N, B, \text{cer}),$$

where

- $N = N_{a_1} \cup N_{a_2} \cup \cdots \cup N_{a_m}$ is a set of nodes such that for each $a \in \{a_1, a_2, \ldots, a_m\}$: $N_a = \{\widehat{a}^{l_1^a}, \widehat{a}^{l_2^a}, \ldots, \widehat{a}^{l_{k_a}^a}\}$,
- $B \subseteq N \times N$ is a set of labeled directed branches such that for any $(\widehat{\phi}^x, \widehat{\phi}^y) \in B$, $\widehat{\phi}^x \in N_{a_{i-1}}$ and $\widehat{\phi}^y \in N_{a_i}$ and $i \in \{2, 3, \ldots, m\}$,
- $\text{cer} : B \to [0, 1]$ is a certainty function labeling branches such that:

$$\text{cer}(\widehat{a}_j^{l_x^{a_j}}, \widehat{a}_k^{l_y^{a_k}}) = \frac{1}{\text{card}(U)} \sum_{u^*\in U^{\mathcal{F}}} f_{\text{inf}}(a_j^{l_x^{a_j}}, u^*) f_{\text{inf}}(a_k^{l_y^{a_k}}, u^*)$$

for any $(\widehat{a}_j^{l_x^{a_j}}, \widehat{a}_k^{l_y^{a_k}}) \in B$.

One can see that we can distinguish particular layers in the set N of nodes of $\mathcal{FFG}(\mathcal{F}(IS))$. The layer N_a, where $a \in \{a_1, a_2, \ldots, a_m\}$, corresponds exactly to one attribute $a \in A$. Each node in the layer N_a corresponds exactly to one linguistic value from the set L^{λ_a} of linguistic values assigned to a linguistic variable λ_a defined for the attribute a. It is worth noting that, in the numerator of the fraction defining the value of the certainty function, the so-called fuzzy cardinality (power) is calculated.

Example 5. For the information system IS^2 and the fuzzified information system $\mathcal{F}(textIS^2)$ considered in Example 2 the fuzzy flow graph corresponding to the information system IS^2 visualized using the Graphviz tool (Ellson *et al.*, 2004) is shown in Fig. 2. ♦

On the basis of a fuzzy flow graph $\mathcal{FFG}(\text{IS})$ corresponding to a given information system IS, we can calculate certainties of non-degenerated sequences over the sets of linguistic values associated with linguistic variables defined for attributes in IS.

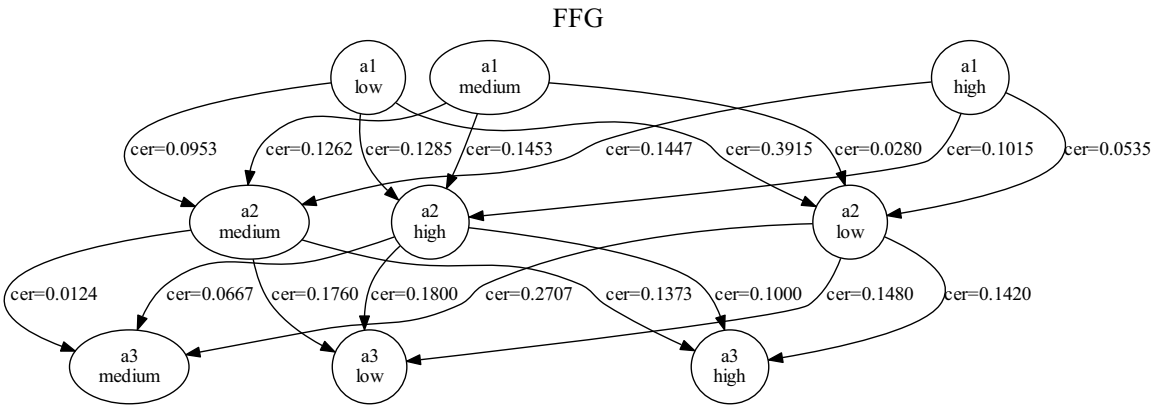

Fig. 2. Fuzzy flow graph corresponding to the information system IS^2.

Definition 10. Let

- $\mathcal{FFG}(\text{IS}) = (N, B, \text{cer})$ be a fuzzy flow graph corresponding to a fuzzified information system $\mathcal{F}(\text{IS}) = (U^{\mathcal{F}}, \Phi, \{V_\phi\}_{\phi\in\Phi}, f_{\text{inf}}^{\mathcal{F}})$ corresponding to an information system $\text{IS} = (U, A, \{V_a\}_{a\in A}, f_{\text{inf}})$, where $A = \langle a_t : t = 1, 2, \ldots, m\rangle$,
- $s(j) = \langle v_{s_1}, v_{s_2}, \ldots, v_{s_k}\rangle$, be a sequence starting at point j, where $j < m$, over the sets of linguistic values associated with linguistic variables defined for attributes from A^2, where $|s(j)| > 1$.

The *certainty* $\text{cer}(s(j))$ of the sequence $s(j)$ is defined as

$$\begin{aligned}&\text{cer}(s(j))\\ &\quad = F((\text{cer}(\widehat{a}_j^{v_{s_1}}, \widehat{a}_{j+1}^{v_{s_2}}), \text{cer}(\widehat{a}_{j+1}^{v_{s_2}}, \widehat{a}_{j+2}^{v_{s_3}}), \ldots,\\ &\qquad \text{cer}(\widehat{a}_{j+k-2}^{v_{s_{k-1}}}, \widehat{a}_{j+k-1}^{v_{s_k}})),\end{aligned}$$

where F is a given triangular norm (see Section 2.4).

Example 6. For the sequences considered in Example 2, one can see that

- $\text{cer}(\widehat{a}_1^{\text{low}}, \widehat{a}_2^{\text{low}}) = 0.3915$,
- $\text{cer}(\widehat{a}_2^{\text{low}}, \widehat{a}_3^{\text{medium}}) = 0.2707$,
- $\text{cer}(\widehat{a}_2^{\text{low}}, \widehat{a}_3^{\text{high}}) = 0.1420$.

Hence, we obtain the certainties of the considered sequences as collected in Table 5. ♦

3. Ant-based clustering for discovering possibly certain sequences

Ant-based clustering is a biologically inspired data clustering technique (Deneubourg *et al*., 1991; Handl *et al*., 2006; Lumer and Faieta, 1994). It belongs to the family of heuristic algorithms used in solving problems with large spaces of possible solutions.

The classic ant-based clustering concerns spaces of cases described by sets of features. In this case, distance measures are commonly used to determine similarities of cases. However, in research, the ant-based approach was also used in clustering other kinds of phenomena, for example, descriptors of decision rules (Parpinelli *et al*., 2002; Pancerz *et al*., 2015). In this section, we propose to use the ant-based clustering procedure in discovering possibly certain sequences created over the sets of attribute values or created over the sets of linguistic values associated with linguistic variables defined for attributes in information systems. The application of a heuristic algorithm (in our proposition, ant-based clustering) is needed since there is a large space of all possible sequences to search.

Let $\text{IS} = (U, A, \{V_a\}_{a\in A}, f_{\text{inf}})$, where $A = \langle a_t : t = 1, 2, \ldots, m\rangle$, be an information system. The number of all possible sequences created over the sets of attribute values can be estimated by

$$\frac{(1+m)m}{2}\prod_{t=1}^{m}|V_{a_t}|,$$

where $|V|$ denotes the cardinality of the set V.

Let $\text{IS} = (U, A, \{V_a\}_{a\in A}, f_{\text{inf}})$, where $A = \langle a_t : t = 1, 2, \ldots, m\rangle$, be an information system and $\{L^{\lambda_a}\}_{a\in A}$ be the family of sets of linguistic values associated with linguistic variables from the family $\{\lambda_a\}_{a\in A}$ defined for attributes from A, where $L^{\lambda_a} = \{l_1^a, l_2^a, \ldots, l_{k_a}^a\}$ for each $a \in A$. The number of all possible sequences created over the sets of linguistic values associated with linguistic variables defined for

Table 4. Certainties of the sequences considered in Example 1.

Triangular norm	$s_1(1) = \langle Z, Z, Y \rangle$	$s_2(1) = \langle X, Y \rangle$
Zadeh's t-norm	0.5000	0.8000
algebraic t-norm	0.3334	0.8000
Lukasiewicz's t-norm	0.1667	0.8000
Einstein's t-norm	0.2857	0.8000
Zadeh's t-conorm	0.6667	0.8000
probabilistic t-conorm	0.8334	0.8000
Lukasiewicz's t-conorm	1.0000	0.8000
Einstein's t-conorm	0.8750	0.8000

Table 5. Certainties of the sequences considered in Example 2.

Triangular norm	$s_1(1) = \langle low, low, medium \rangle$	$s_2(2) = \langle low, high \rangle$
Zadeh's t-norm	0.2707	0.1420
algebraic t-norm	0.1060	0.1420
Lukasiewicz's t-norm	0.0000	0.1420
Einstein's t-norm	0.0734	0.1420
Zadeh's t-conorm	0.3915	0.1420
probabilistic t-conorm	0.5562	0.1420
Lukasiewicz's t-conorm	0.6622	0.1420
Einstein's t-conorm	0.5987	0.1420

attributes can be estimated by

$$\frac{(1+m)m}{2} \prod_{t=1}^{m} |k_{a_t}|.$$

One can see that the search of the entire space of sequences leads to the exponential time complexity with respect to the number m of attributes in a given information system.

Let $\mathcal{FG}(IS) = (N, B, \text{cer})$ be a flow graph that is either a rough set flow graph $\mathcal{RSFG}(\text{IS}) = (N, B, \text{cer}, \text{str}, \text{cov})$ without a strength function 'str' and a covering function 'cov' or a fuzzy flow graph $\mathcal{FFG}(\mathcal{F}(\text{IS})) = (N, B, \text{cer})$. The main idea of the ant-based clustering procedure is to concatenate two adjacent sequences to make a new longer sequence. The probability of concatenation depends on the certainty of the new sequence.

Definition 11. Let

- IS $= (U, A, \{V_a\}_{a \in A}, f_{\text{inf}})$, in which $A = \langle a_t : t = 1, 2, \ldots, m \rangle$ is an information system,
- $s(j)$ be a sequence starting at point j, where $j \leq m$, over the sets of attribute values in IS,
- $s'(j')$ be a sequence starting at point j', where $j' \leq m$, over the sets of attribute values in IS.

The sequences $s(j)$ and $s'(j')$ are *adjacent* if and only if either $j + |s(j)| = j'$ or $j' + |s'(j')| = j$.

A formal description of the algorithm for ant-based clustering for discovering possibly certain sequences is presented in Algorithm 3. Possibly certain sequences are sequences with certainties as high as possible. The algorithm has polynomial time complexity.

In Algorithm 3, the following functions are used:

- $selectRandomlySequence$, the function selecting randomly one sequence from the set of sequences,
- $selectRandomlyNumberFromUnitInterval$, the function selecting randomly one real number from the unit interval $[0.0, 1.0]$,
- $concatenate$, the function concatenating two adjacent sequences to make a new longer sequence.
- $certainty$, the function calculating the certainty of the given sequence (see Definitions 8 and 10).

Moreover, each ant object stores information on the carried sequence ($carriedSequence$) by the ant and each sequence object stores information on whether the sequence is carried ($isCarried$) by the ant.

The set of possibly certain sequences discovered by Algorithm 3 can be filtered. We can, for example,

- remove degenerated sequences,

Algorithm 1. Ant-based clustering for discovering possibly certain sequences.

Require: $\mathcal{FG}(\text{IS}) = (N, B, \text{cer})$: the flow graph, n: the number of iterations performed for the clustering process, $ants$: the set of ants

$sequences \leftarrow \emptyset$;
for each node $n^x \in N$ **do**
  $sequence \leftarrow \{x\}$;
  $sequence.isCarried \leftarrow$ false;
  $sequences \leftarrow sequences \cup \{sequence\}$;
end for
for each $ant \in ants$ **do**
  $sequence \leftarrow selectRandomlySequence(sequences)$;
  if $sequence.isCarried =$ FALSE **then**
    $sequence.isCarried \leftarrow$ TRUE;
    $ant.carriedSequence \leftarrow sequence$;
  else
    $ant.carriedSequence \leftarrow$ NULL;
  end if
end for
for $k \leftarrow 1 \ldots n$ **do**
  for each $ant \in ants$ **do**
    if $ant.carriedSequence =$ NULL **then**
      $sequence \leftarrow selectRandomlySequence(sequences)$;
      if $sequence.isCarried =$ FALSE **then**
        $sequence.isCarried \leftarrow$ TRUE;
        $ant.carriedSequence \leftarrow$ **sequence**;
      end if
    else
      $sequence \leftarrow selectRandomlySequence(sequences)$;
      if $sequence.isCarried =$ FALSE **then**
        if $areAdjacent(sequence, ant.carriedSequence) =$ TRUE **then**
          $newSequence \leftarrow concatenate(sequence, ant.carriedSequence)$;
          if $selectRandomlyNumberFromUnitInterval \leq certainty(newSequence)$ **then**
            $newSequence.isCarried \leftarrow$ FALSE;
            $sequences \leftarrow sequences \cup \{newSequence\}$;
            $ant.carriedSequence \leftarrow$ NULL;
          end if
        end if
      end if
    end if
  end for
end for

- remove sequences being subsequences of other sequences,
- remove sequences with certainties below a given threshold,
- remove sequences of the length below a given threshold.

It is worth noting that, in the case of rough set flow graphs, we can also discover possibly strong sequences replacing the certainties of sequences by the strengths of sequences.

We have tested the presented approach on MMPI data consisting of the so-called profiles of patients screened with the MMPI (Minnesota Multiphasic Personality Inventory) standardized psychometric test (Nichols, 2011). The profile is composed of an ordered values (generally, between 0 and 120) of thirteen scales. The data set used in experiments was collected for research by in a psychological outpatient clinic (Duch *et al*., 1999). It includes profiles of 1710 women. For each scale, a set of ten linguistic values was used to fuzzify its numerical values. The boundary linguistic values were described by the trapezoidal shaped membership

Table 6. Results of experiments for the algebraic t-norm.

Fold ID	No. of sequences found	Avg. actual certainty	Avg. difference	Std dev.
1	46	0.0395	0.0090	0.0099
2	44	0.0232	0.0041	0.0044
3	42	0.0260	0.0048	0.0050
4	42	0.0331	0.0064	0.0075
5	48	0.0275	0.0056	0.0054
6	47	0.0247	0.0068	0.0071
7	37	0.0215	0.0050	0.0067
8	46	0.0372	0.0076	0.0077
9	45	0.0287	0.0084	0.0089
10	42	0.0408	0.0055	0.0055

Table 7. Results of experiments for the probabilistic t-conorm.

Fold ID	No. of sequences found	Avg. actual certainty	Avg. difference	Std dev.
1	54	0.2868	0.0204	0.0166
2	63	0.2950	0.0331	0.0202
3	55	0.2433	0.0063	0.0059
4	61	0.3187	0.0158	0.0121
5	54	0.3317	0.0244	0.0134
6	54	0.2264	0.0188	0.0133
7	68	0.2493	0.0104	0.0086
8	56	0.3132	0.0110	0.0080
9	63	0.3193	0.0117	0.0089
10	62	0.3969	0.0092	0.0065

functions whereas the remaining ones by the triangular shaped membership functions. The experiments were performed using the ten-fold cross validation approach. In each iteration, nine parts were used in the training stage to create a fuzzy flow graph and to generate possibly certain sequences of linguistic values using the ant-based clustering. One part was used to determine actual certainties of the sequences found in the training stage. Next, we calculated the average value of actual certainties of the sequences, the average value and the standard deviation for differences between the actual certainties of the sequences and the predicted certainties (determined in the training step) of the sequences. Moreover, several triangular norms were used to determine certainties of sequences. In experiments, we used a tool called CLAPSS (Classification and Prediction Software System) (Pancerz, 2015), where the proposed approach has been implemented.

The results of experiments are collected in Tables 6 and 7 for the algebraic t-norm and the probabilistic t-conorm, respectively. In each fold, several dozen sequences were found by the ant-based clustering. One of the possibly certain sequences found has form

$$L = AVG_3 \rightarrow F = AVG_3 \rightarrow K = AVG_2 \rightarrow 1.Hp = AVG_3 \rightarrow 2.D = AVG_1 \rightarrow 3.Hy = HIGH \rightarrow 4.Ps = AVG_3 \rightarrow 5.Mk = AVG_2 \rightarrow 6.Pa = AVG_3,$$

where $L, F, K, \ldots$ are scales (Nichols, 2011) and $AVG_3, AVG_2, \ldots$ are linguistic values used in the fuzzification process (Pancerz *et al.*, 2018). The average aggregated (using triangular norms) certainties for sequences found were compared with those predicted on the basis of a flow graph. On the basis of differences, the errors of prediction were determined.

In general, the experiments showed that the prediction error of certainties of sequences found is about 10%. Application of t-norms for determining certainties of sequences may be treated as some pessimistic approach, whereas application of t-conorms may be identified as some optimistic approach.

4. Discussion and conclusions

Flow graphs (both those based on rough set theory and those based on fuzzy set theory) are very useful tool for modeling sequence data. Their potential in machine learning seems not to be fully exploited (one can see

a relatively low number of publications in this area). Therefore, we have proposed an approach to sequence data mining based on fuzzy flow graphs. Flow graphs have become a base model of data flow. On the basis of this model, searching for possibly certain sequences appearing in data has been performed. Owing to the large space of all possible sequences, we have proposed to use an ant-based clustering procedure that represents one of the heuristic approaches.

Experiments demonstrate that the proposed approach is promising in solving sequence data mining problems. It can be used in various problems concerning prediction of sequence appearance in the future. The certainty of such events is determined not on the basis of training sequences (ordered sets of elements) treated as a whole, but on the basis of aggregated certainties of coexistences of consecutive adjacent elements in sequences. We can say that local coexistences are taken into consideration. The knowledge of certainties of coexistences is extracted on the basis of flow graphs modeling sequence data. This is a new idea, different from those based on global coexistences of all elements of sequences (discovering frequent item sets is an example). Due to the heuristic character of the proposed approach, large spaces of sequences can be mined. Because of application of the ant-based clustering, the proposed approach requires experimental selection of parameters (the number of ants, the number of iterations, etc.) as well as functions (dropping down and picking up functions, t-norms and t-conorms). In some cases, this can be done using a trial-and-error method. This is the main disadvantage of the proposed approach.

Our further research will be focused mainly on two directions. Firstly, in the case of fuzzy flow graphs, we can test a variety of shapes of membership functions used to model linguistic values as well as a variety of triangular norms used to determine certainties of sequences found. Secondly, a challenging thing is to use approaches in which the domain knowledge is taken into consideration in graph mining (Bazan *et al*., 2013; Pancerz, 2016).

References

Bazan, J.G., Buregwa-Czuma, S. and Jankowski, A.W. (2013). A domain knowledge as a tool for improving classifiers, *Fundamenta Informaticae* **127**(1–4): 495–511.

Deneubourg, J., Goss, S., Franks, N., Sendova-Franks, A., Detrain, C. and Chrétien, L. (1991). The dynamics of collective sorting: Robot-like ants and ant-like robots, *Proceedings of the 1st International Conference on Simulation of Adaptive Behaviour: From Animals to Animats 1, Paris, France*, pp. 356–365.

Dong, G. and Pei, J. (2007). *Sequence Data Mining*, Springer-Verlag, New York, NY.

Duch, W., Kucharski, T., Gomuła, J. and Adamczak, R. (1999). *Machine Learning Methods in Analysis of Psychometric Data: Application to Multiphasic Personality Inventory MMPI-WISKAD*, CMK, Toruń, (in Polish).

Ellson, J., Gansner, E.R., Koutsofios, E., North, S.C. and Woodhull, G. (2004). Graphviz and Dynagraph—Static and dynamic graph drawing tools, *in* M. Jünger and P. Mutzel (Eds), *Graph Drawing Software*, Springer, Berlin/Heidelberg, pp. 127–148.

Ford, L.R. and Fulkerson, D. (2010). *Flows in Networks*, Princeton University Press, Princeton, NJ.

Handl, J., Knowles, J. and Dorigo, M. (2006). Ant-based clustering and topographic mapping, *Artificial Life* **12**(1): 35–62.

Huang, K.-Y. and Chang, C.-H. (2008). Efficient mining of frequent episodes from complex sequences, *Information Systems* **33**(1): 96–114.

Klement, E.P., Mesiar, R. and Pap, E. (2000). *Triangular Norms*, Springer, Dordrecht.

Kumar, P., Kumar, P., Krishna, P.R. and Raju, S.B. (2011). *Pattern Discovery Using Sequence Data Mining: Applications and Studies*, IGI Global, Hershey, PA.

Lumer, E. and Faieta, B. (1994). Diversity and adaptation in populations of clustering ants, *Proceedings of the 3rd International Conference on Simulation of Adaptive Behaviour: From Animals to Animats 3, Brighton, UK*, pp. 501–508.

Mannila, H., Toivonen, H. and Verkamo, A. (1997). Discovering frequent episodes in event sequences, *Data Mining and Knowledge Discovery* **1**(3): 259–289.

Marwala, T. (2013). *Introduction to Economic Modeling*, Springer, London.

Mieszkowicz-Rolka, A. and Rolka, L. (2006). Flow graphs and decision tables with fuzzy attributes, *in* L. Rutkowski *et al.* (Eds), *Artificial Intelligence and Soft Computing (ICAISC 2006)*, Springer, Berlin/Heidelberg, pp. 268–277.

Mitsa, T. (2010). *Temporal Data Mining*, CRC Press, Boca Raton, FL.

Nichols, D. (2011). *Essentials of MMPI-2 Assessment*, John Wiley, Hoboken, NJ.

Pancerz, K. (2015). On selected functionality of the classification and prediction software system (CLAPSS), *Proceedings of the International Conference on Information and Digital Technologies (IDT'2015), Zilina, Slovakia*, pp. 278–285.

Pancerz, K. (2016). Paradigmatic and syntagmatic relations in information systems over ontological graphs, *Fundamenta Informaticae* **148**(1–2): 229–242.

Pancerz, K., Lewicki, A. and Tadeusiewicz, R. (2015). Ant-based extraction of rules in simple decision systems over ontological graphs, *International Journal of Applied Mathematics and Computer Science* **25**(2): 377–387, DOI: 10.1515/amcs-2015-0029.

Pancerz, K., Lewicki, A., Tadeusiewicz, R. and Warchoł, J. (2013). Ant-based clustering in delta episode information systems based on temporal rough set flow graphs, *Fundamenta Informaticae* **128**(1–2): 143–158.

Pancerz, K., Paja, W., Sarzyński, J. and Gomuła, J. (2018). Determining importance of ranges of MMPI scales using fuzzification and relevant attribute selection, *Procedia Computer Science* **126**: 2065–2074.

Pancerz, K., Paja, W., Wrzesień, M. and Warchoł, J. (2012). Classification of voice signals through mining unique episodes in temporal information systems: A rough set approach, *Proceedings of the 21st International Workshop on Concurrency, Specification and Programming (CS&P 2012), Vilamoura, Algarve, Portugal*, pp. 280–291.

Parpinelli, R.S., Lopes, H.S. and Freitas, A.A. (2002). An ant colony algorithm for classification rule discovery, *in* H.A. Abbass *et al.* (Eds), *Data Mining: A Heuristic Approach*, IGI Global, Hershey, PA, pp. 191–208.

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

Pawlak, Z. (2005). Flow graphs and data mining, *in* J. Peters and A. Skowron (Eds), *Transactions on Rough Sets III*, Springer-Verlag, Berlin/Heidelberg, pp. 1–36.

Tadeusiewicz, R. (2015). Neural networks as a tool for modeling of biological systems, *Bio-Algorithms and Med-Systems* **11**(3): 135–144.

Arkadiusz Lewicki has a PhD in computer science. He is with the Department of Applied Informatics of the University of Information Technology and Management in Rzeszów. His research focuses mainly on swarm intelligence, neural networks and evolutionary computation. His current research interests also include metaheuristics and local search methods for combinatorial optimization and parallel and distributed computing.

Krzysztof Pancerz is an associate professor of computer science at the University of Rzeszów. He received his MSc in electrical engineering, in 1998 from the Rzeszów University of Technology, and both his PhD and DSc (habilitation) in computer science from the Institute of Computer Science, Polish Academy of Sciences, in 2006 and 2018, respectively. His research interests concern computational intelligence, knowledge engineering, unconventional computing, and computer-aided diagnosis. He has published over 100 research papers in international journals, monographs and conference proceedings.

Received: 21 November 2019
Revised: 6 March 2020
Re-revised: 4 June 2020
Accepted: 8 June 2020