HUMAN TECHNOLOGY

ISSN: 1795-6889

https://ht.csr-pub.eu Volume 17(3), December 2021, 213–234

THE TEMPORAL INFERENCE WITH THE USE OF ANT-BASED CLUSTERING ALGORITHM AND FLOW GRAPHS IN THE PROBLEM OF PROGNOSING COMPLICATIONS OF MEDICAL SURGICAL PROCEDURES

Arkadiusz Lewicki
Faculty of Applied Computer Science
University of Information Technology
and Management in Rzeszow
Poland

Krzysztof Pancerz
Institute of Philosophy
John Paul II
Catholic University of Lublin
Poland

Leszek Puzio
Faculty of Applied Computer Science
University of Information Technology
and Management in Rzeszow
Poland

Abstract: *In the era of a rapidly aging European society, the demand for proven clinical decision support systems, links health observations with medical knowledge in order to assist clinicians in decision making is constantly growing. An increasing problem for this type of systems is not only the size of the processed data sets but also the heterogeneity of these data. Clinical forecasting often requires processing of both numerical data and multi-category data which are temporal. The conducted research has shown that a good solution to this problem may lie in the use of temporal inference, the ant-based clustering algorithm, rough sets, and fuzzy sets. The experiments used a real set of medical data representing cases of a disease that significantly reduces a woman's quality of life. Each case of uterine myoma disease (which affects more than 50% of women over the age of 35) is represented by more than 140 heterogeneous features. An incorrect decision about the type of surgery (thermoablation or surgery) not only affects female fertility but also the high risk of complications. Therefore, the solution discussed in this paper may turn out to be extremely important.*

Keywords: *temporal inference, flow graphs, rough sets, fuzzy sets, ant-based clustering.*

DOI: https://doi.org/10.14254/1795-6889.2021.17-3.3

INTRODUCTION

The fast growth in the amount of available information collected in an electronic form and the development of digital data acquisition and recording technologies means that the techniques of exploring and discovering knowledge in data are used in an increasing number of fields. However, this requires an interdisciplinary approach. The domain data must be properly understood before the essential features of the available data set can be selected. Only then we can start coding, normalizing, exploring and analysing data. The data mining becomes a key tool for the competitive advantage of companies in terms of discovering patterns and dependencies, constructing analytical models, assessing the degree to which models fit the data and interpreting the results. It also revolutionizes all areas related to the quality of human life, primarily in terms of protection and improvement of human health. Therefore, in medicine, the demand for data analysis tools that use automatic and semi-automatic methods to identify and describe patterns, trends, or relationships in data is growing more than ever. However, this is associated with many difficulties. The basic problem is that the collected data resources have many incomplete features. The data are heterogeneous because they include not only numerical values (integer and real values) but also categorical values and multi-categories. In addition, the data require the application of various fields of data science due to the need to process multidimensional data structures, including irregularly sampled processes with space-time dependencies. The results of the research on available medical databases, which were recently published in *The Lancet Digital Health* (Wen et al., 2021) clearly show the problem of the lack of complete data. According to the authors, machine learning algorithms, that are to help in the early detection of dangerous diseases (e.g. skin cancer), learn from incomplete databases. Consequently, their effectiveness cannot be guaranteed. This has also been confirmed in other studies (Daneshjou, Smith & Sun, 2021; Pawlowski, 2019). In addition, the available medical data resources include multidimensional heterogeneous data structures (in the form of time series, numerical data, categorical data and image files). It is a significant challenge in terms of correct data preprocessing, so that the processed data resources can be used effectively (Thirumahal, Sadasivam, 2020; Ren, Lu, Wang, 2018). The correct approach to machine learning requires the analysis of data structure and relationships, development or selection of an appropriate model of inference, as well as visualization and interpretation of the obtained results. Therefore, the analysis and inference are carried out using both quantitative and qualitative methods. This approach takes into account linear, nonlinear, discrete and continuous models, including data reduction algorithms, feature extraction, data discretization, decision rule generation, neural network models, frequency analysis models, and many statistical analysis tools and models. However, most of them work well when dealing with a large number of samples, a small amount of missing data and multi-criteria data. In any other case (Salcedo-Bernal, Villamil-Giraldo, Moreno-Barbosa, 2016; Dhillon & Singh, 2019), a different method is used that will allow the achievement of satisfactory results (relating to an acceptable error value at the output). These include solutions using fuzzy sets (Vlamou, Papadopoulos, 2019) or rough sets (Burney & Abbas, 2015). They allow inference based on heterogeneous data sets, but require complete data sets. Therefore, in the opinion of the authors of the article, the appropriate approach to the problem may be the use of an ant-based clustering algorithm and flow graphs. The ant-based clustering algorithm is an example of the algorithm that does not require specifying the expected number of clusters. This is a significant advantage as compared

to the most frequently used: K-means algorithm, hierarchical algorithms or Gaussian Mixture algorithm (Su, Dy, 2007). The approach presented in the article was tested on a real medical data set.

The analytical process of knowledge discovery to support decision-making always requires proper data preparation (Tripathi, Muhr, Brunner, Jodlbauer, Dehmer & Emmert-Streib, 2021; Alnoukari & El Sheikh, 2012). Therefore, data validation, cleansing, re-encoding and variable selection are performed. The data encoding process includes transformation, cleansing and data quality assessment. Feature fields with missing data must be properly completed or marked. The next step is to select the features for analysis, because not all variables will be meaningful or valuable in the proposed solution. In addition, all data should be normalized. A data set prepared in such a way can be processed by an appropriate decision inference algorithm. This is shown in Figure 1.

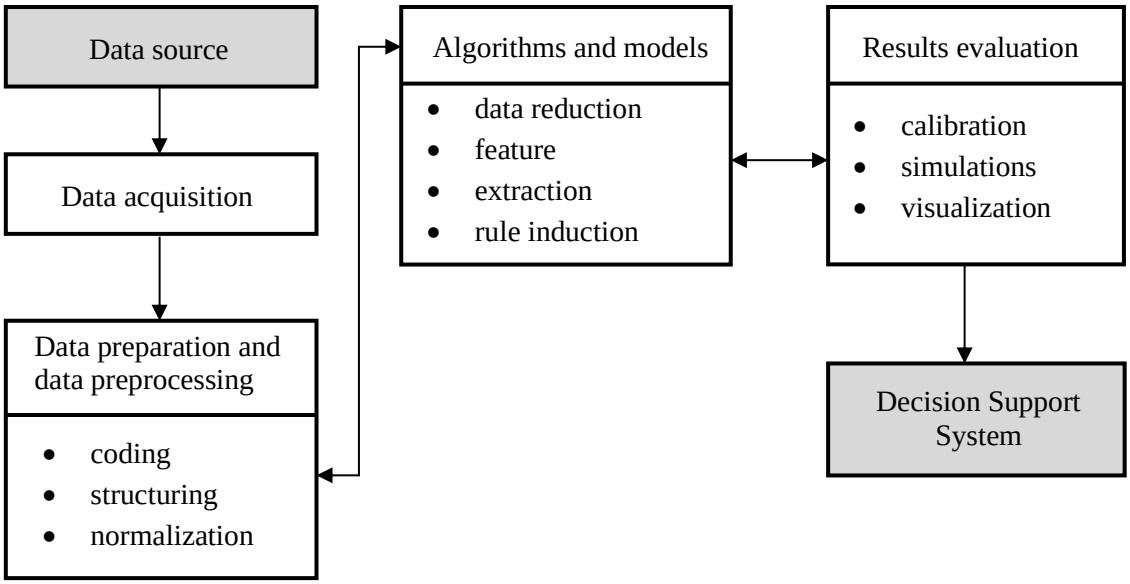

Figure 1. Data processing model for forecasting purposes.

The decision support system based on the mechanisms of relational analysis of the aggregated features of the available data set is a solution expected in many fields of medicine. One of them is gynaecology. No proven solution to the problem of predicting the effectiveness of thermoablation of fibroids treatment has been developed so far in this medical field. Currently, the qualification of the patient for the treatment of thermal ablation of fibroids is based on a gynecologist's consultation with a radiologist. However, the research conducted has shown (Lozinski, Filipowska, Gurynowicz, Pyka, Ciebiera, 2020) that it is not always the right decision. Uterine fibroids are benign tumours of the female reproductive organs (De La Cruz & Buchanan, 2017). They affect even every fourth woman of childbearing age. The patient is usually referred for removal of the uterine body or the entire uterus in the event of heavy bleeding, abdominal pain and anaemia. However, it is associated with the loss of female fertility. The ultrasonic thermoablation treatment is a hope for its preservation. The ultrasonic thermoablation is a non-invasive procedure (Kalamarz, Zagrobelna & Pyziak, 2017) that has only been used in gynaecology for several years. It uses a 3-Tesla device. The energy of the

focused ultrasound beam generated by this device is precisely directed from ultrasound or magnetic resonance imaging. The high-frequency ultrasound wave causes a temperature increase in the range of 60-100 degrees in the target tissue, protein denaturation and tumour degradation as a result. However, the ultrasonic thermoablation treatment is very expensive. The decision about the positive qualification of the patient for the procedure of ablation is now made by doctors on the basis of 4 features selected from a set of over 140 features representing the detailed results of diagnostic tests. The analysis usually concerns only such features as: the location of the myoma, its relation to the environment, the thickness of the adipose tissue and the type of myoma according to the Funaki classification (Sainio, Saunavaara, Komar, Mattila, Otonkoski & Joronen, 2021). However, the available literature (Lozinski, Filipowska, Gurynowicz, Pyka & Ciebiera, 2020; Verpalen, Anneveldt, Nijholt, Schutte, Dijkstra, Franx, Bartels, Moonen, Edens & Boomsma, 2019) and the analysis of the available control results of 218 patients after 6 and 12 months show that this approach is only associated with an accuracy of 25% to 35% (Lozinski, Filipowska, Gurynowicz, Pyka, Ciebiera, 2020). Therefore, we have created and next evaluated an IT system of decision support for the problem defined above.

The proposed approach is related to the analysis of data containing information obtained before the procedure and 3 and 6 months after the thermoablation procedure. The data collection procedure results from the standard medical procedure adopted by the physicians in the considered medical problem. It is possible to achieve valuable knowledge by building a flow graph in which nodes represent specific clusters determined on the basis of features related to a given temporal information. The use of data collected by physicians does not give satisfactory results in qualifying patients for thermoablation. Hence, there is a need to use machine learning tools to extract knowledge from data. Each node represents a specific time event and information about the affiliation of each processed data vector to a specific group of objects with similar characteristics. Therefore, data sets related to significant temporal features must be clustered in advance in order to create a flow graph. This is due to the fact that vectors are created in different feature spaces describing patients at distinguished moments in time. Which features appear at each stage depends on the medical procedure determined by physicians. Therefore, there is no one feature space considered in all stages. The rationale behind the clustering and flow graph approach is the desire to link two perspectives in the inference problem under consideration: a spatial perspective (clustering in feature spaces) and a temporal perspective (information flow over time represented by a flow graph). The metaheuristic ant-based algorithm was used due to multidimensional data structures, missing data and the inability to indicate the number of result groups for this purpose. The authors of the article did not find other solutions allowing for effective inference from incomplete multidimensional data.

The data set used in our study was obtained thanks to cooperation with a medical clinic. It contains 145 heterogeneous features (including integer, real, categorical and multi-category values) and 218 data records. Each of the 218 data vectors represents a separate patient with severe symptoms of uterine fibroids. The data set was initially processed and then divided into 3 subsets (with different features), representing respectively the test results necessary to make a decision about thermoablation treatment and its implementation, the results of control tests performed 3 months after the procedure and the results of control tests performed 6 months after the procedure. The data feature interdependencies are determined by physicians on the

basis of adopted medical procedures and the medical knowledge available. The first of the processed data subsets is the data set of medical test results before the thermal ablation procedure. These data include blood counts, ultrasound, detailed medical history, history of diseases, history of pharmacological treatment, measurements of the size of each detected myoma and detailed gynecological diagnostics. This set consists of 91 unique features. The second collection concerns the results of the patient's control test, 3 months after the procedure. It consists of 29 different features from those in the first data set. These data include the results of ultrasonography performed 3 months after the thermoablation treatment, post-treatment NPV and magnetic resonance imaging. The last subset of data processed is a set of another 25 unique features. They represent the patient's control test results 6 months after the thermal ablation procedure. The physicians define the range of the patient's control test at each stage. Therefore, there does not exist any fixed relationship between features from particular collections. There is only a relationship resulting from the physician's decision in regard to the scope of the tests. All non-numeric feature values (including categorical values) in the above data sets have been encoded as numerical data. The one-hot encoding method was used for this purpose and then the data was normalized. One-hot encoding is a process in data processing that is applied to categorical data in order to convert it into a binary vector representation. Individual values of such a vector can then be represented as a sequence of features. An exemplary (1 of 218) input data vector of a data subset is shown below:
[1.0;1.0;1.0;0.27586;0.4;0.96753;0.0;0.0;0.0;0.0;0.0;0.0;0.058823;1.0;0,0;0.0;0.0;1.0;1.0;1.0;
1.0;0.0;1.0;0.0;0.0;0.0;1.0;0.0;0.0;0.0;0.0;0.0;0.0;0.0;1.0;1.0;0.00151;0.68582;0.67677;0.480
78;0.21003;0.027399;0.362076;0.009971;0.17734;0.01397;0.350981;0.00491;0.33333;0.626
16;0.55652;0.56140;0.27447;0.6875;0.0;0.0;0.8;0.24137;0.0;1.0;0.84347;0.62385;0.60629;0.
40255;0.0;0.0;0.0;0.0;0.0;0.0;1.0;0.0;0.0;1.0;0.0;0.0;0.0;0.46341;0.0;0.03678;0.23990;0.4243
3;0.325;0.23529;0.27234;0.10455;0.64285;0.44;1.0;1.0;0.42307;2.0]
The subsets were clustered and a flow temporal inference graph was built. There are many algorithms for grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups (Singh, Srivastava, 2020). The division can be made according to data types (e.g. numbers, text data, images, etc.), the mechanism of generating clusters (deterministic and probabilistic algorithms) or the method of obtaining clusters (hierarchical and flat algorithms). Unfortunately, many traditional clustering algorithms have many disadvantages. They are primarily computationally inefficient because they work well for a small number of features of grouped data. They are also sensitive to outliers. At this point, it is important to refer to (Handl, Knowles, Dorigo, 2003), where the authors discuss the advantages of using the ant-based clustering algorithm over other clustering algorithms. Among other things, the authors draw attention to the following facts: (i) the nature of the ant-based algorithm makes it fairly robust to the effects of outliers within the data, (ii) ant-based clustering has the capacity to work with any kind of data that can be described in terms of symmetric dissimilarities, and it imposes no assumption on the shape of the clusters it works with. Moreover, other clustering algorithms often also require a priori information about the number of expected clusters. However, heuristic algorithms deal with these problems well. One of them is the ant-based clustering algorithm (Boryczka, 2008; Pancerz, Lewicki, 2012).

The flow graph is a suitable tool to model information flow in case of heterogeneous domains. As it will be shown in the next sections, we cannot treat our problem in terms of time series analysis, because we do not deal with the same quantities changing in time.

METHODOLOGY

Algorithm of Ant Collective Intelligence in the Task of Data Clustering

The ant-based clustering algorithm (Boryczka, 2008; Pancerz, Lewicki, 2012) is a population-based algorithm in which subsequent generations of virtual ant are looking for good quality solutions. It was created on the basis of observation of the ability of ants to stack the bodies of dead ants (Deneubourg, Goss, Franks, Sendova-Franks, Detrain & Chretien, 1991). Its basic operating principle relates to the ability of the agent (ant) to pick up, transport and drop items. The possibility of picking up and dropping an element is determined by the probability. The value of the probability depends on the neighbouring elements in the available search space. Ants prefer to pick up objects that are isolated or adjacent to different objects. Moreover, they tend to drop objects that are in the vicinity of similar objects. Objects are grouped within the explored area as a result of this approach. The Ant-based clustering algorithm does not use the phenomenon of leaving a pheromone trace. Moreover, the size of the ant (agent) population does not have a large impact on the algorithm's performance.

The ant-based clustering algorithm was used first to group the first set of objects, then to group the second set of objects, and after that to the third set. The first data set was a set of patient test results prior to the uterine myoma ablation procedure. This set consists of 218 data vectors. Each vector represents a different patient and it is described by 91 unique features. Each of the values of the processed features (after carrying out the coding and normalization process) was a numerical value. Each vector of this set consists of the same number of features, which give information on: past thermoablation, operations, contraindications (that we know), patient's age, weight, height, number of births, number of miscarriages, obstruction of the fallopian tube, anovulation, partner's sperm, date of first myoma detection, heavy menstruation, anemia, HB before the procedure, intermenstrual bleeding, abdominal pain, painful menstruation, painful intercourse, pressure on the bladder, pressure on the rectum, attempts to treat fibroids or fibroids, as well as laparoscopy data, ultrasound data, detailed information on the size of each detected myoma and information about its position. The second set of data was the results of the patients' control tests performed 3 months after the thermoablation procedure. Data vectors of this set were represented by attributes corresponding to such features as: change in the volume of the myoma, position in relation to the uterus, position in relation to the abdominal cavity, length of the myoma, width of the myoma, volume of the myoma visualized using ultrasound, menstrual bleeding, bleeding between periods, soreness of menstruation, the patient's well-being, NPV, quality of life, Funaki classification of myoma size change, the degree of hardness of the lower abdomen, evaluation of the pressure on the bladder, parameters of ultrasound examination with contrast and MRI parameters. This set was also converted into numerical values after the coding and normalization process. These features (apart from the contrast ultrasound) also represented the third data set. The third set of data was the results of the patients' control tests performed 6 months after the thermoablation procedure. Each of the three sets listed above consisted of 218 data vectors. The expected result of grouping each object of the first, second and third data set was the information about the cluster number in which the object was placed. In general, there is no requirement that the feature sets defined at each stage are related to each other. The physicians define the range of the examination at each stage. In the extreme case, the patient may be described by a completely different set of features

at each stage. The only determinant is the fact that these features allow physicians to determine the condition of the patient at a given time for the selected disease entity.

Objects representing the input data vectors (recordset) were randomly placed in space (in the form of a toroidal grid) in the initialization phase. The number of agents (ants) was set to the same value as the number of data vectors processed (218). Each agent has been associated with a different object and placed on the grid at random coordinates. The next step was to iterate for a given value. Experiments showed that the best results were obtained for a value equal to 500. Each agent moved randomly around the grid picking and dropping the data objects. The decision of dropping an object depended on the calculated probability value. The probability of dropping an object was given by a formula:

$$P_d(o_i) = \begin{cases} 1, & if\ f(i) \geq 1 \\ \frac{1}{f(i)^{4.}}, & else \end{cases}$$

where:

- P_d (o_i) is the probability of dropping object o_i,
- f(i) is a modified version of Lumer and Faieta's neighbourhood function (Lumer, Faieta, 1994).

The neighbourhood function f(i) was given by formula:

$$f(i) = \begin{cases} \frac{1}{\delta^2} \sum_j \left[1 - \frac{d(i,j)}{\alpha}\right], if\ f > 0\ and\ \left(1 - \frac{d(i,j)}{\alpha}\right) > 0 \\ 0, \qquad otherwise \end{cases}$$

where:

- d(i, j) is the dissimilarity (Euclidean distance) between a pair of objects (o_i, o_j),
- α is a factor that defines the scale for dissimilarity,
- δ^2 is a neighbourhood size ($\delta^2 \in \{9;25\}$).

A modification of the Lumer and Faieta algorithm was applied here in order to improve convergence due to the fact that the problem under consideration concerns the grouping of data in the context of knowledge discovery. The algorithm without a modified version of Lumer and Faieta's neighbourhood function was less efficient (Pancerz, Lewicki, 2012).

The similarity function was computed on the basis of the Euclidean distance. The agent left the object where it currently was on the grid in case of decision to drop an element. He searched for a new object to pick up at the next step. It was related to the probability function defined by formula:

$$P_p(o_i) = \begin{cases} 1, & if\ f(i) > 1 \\ \frac{1}{f(i)^{2.}}, & otherwise \end{cases}$$

where:

- Pd (o_i) is the probability of picking object o_i,

– f(i) is a modified version of Lumer and Faieta's neighbourhood function.

8 unambiguously distinguishable clusters were obtained for each grouped data set. The number of clusters is determined automatically, without any interference in the analysis. They were numbered from 0 to 7. The process of grouping the first data set gave the following result:
[2;1;1;2;1;2;1;1;5;1;1;2;4;1;2;1;6;5;5;1;1;5;2;2;2;2;2;5;1;1;5;5;2;1;1;2;5;2;2;4;1;1;2;2;2;2;4;2
;2;1;2;2;2;2;5;1;2;1;2;2;4;4;2;2;1;2;4;1;2;5;2;2;2;2;5;4;5;2;1;1;2;5;2;1;4;2;2;6;4;4;1;4;4;2;5;2
;1;1;1;6;0;1;1;2;2;4;4;2;4;2;5;5;1;1;1;4;4;2;1;1;5;2;1;0;2;4;2;6;6;6;4;1;2;6;2;6;2;6;1;2;6;4;6;6
;6;0;0;6;0;0;7;0;0;0;0;0;0;7;5;0;7;7;0;0;6;0;0;7;7;0;0;0;0;0;0;0;7;7;3;7;7;7;7;3;7;3;3;3;3;3;3;3;3
;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3;3].

It means that the first object of first data set belongs to the cluster number 2, the second object belongs to the cluster number 1, etc. They are values of features at the first time point of the flow graph presented in this article.

The process of grouping the second data set gave the following result:
[4;4;7;1;7;4;2;2;7;5;1;1;5;4;1;1;6;7;4;7;5;5;7;2;4;1;1;5;1;5;7;7;1;1;1;1;7;1;7;5;2;1;2;5;4;1;5;5
;2;2;1;1;5;1;4;1;1;5;1;1;7;1;1;5;0;5;1;2;1;1;1;1;7;5;7;2;5;1;5;1;7;1;1;5;2;1;5;6;2;7;0;2;1;5;1;1
;1;5;2;4;3;7;5;1;1;7;1;5;1;1;1;7;7;1;0;1;1;0;6;2;1;1;1;6;7;2;7;6;6;5;1;2;0;6;1;6;7;4;7;4;4;5;5;5
;6;5;5;6;3;0;0;0;3;3;6;2;6;0;3;6;2;6;0;4;0;6;3;4;3;0;0;3;3;0;6;3;0;3;0;4;0;0;0;0;0;0;0;3;0;3;0;3
;0;0;0;0;0;3;0;3;0;3;0;3;0;3;0;3;0;0;0;0;3;0;3;3;0;6;0;3].

They are values of features at the second time point of the flow graph.

The process of grouping the third data set gave the following result:
[2;2;1;3;7;2;3;7;3;1;2;2;3;2;3;3;4;7;2;1;1;7;2;7;2;3;2;1;3;2;3;1;3;1;1;3;2;1;2;1;1;3;1;3;2;3;2;1
;7;2;3;3;3;1;2;7;3;3;3;3;7;3;3;1;3;1;3;7;2;3;3;3;1;2;1;3;1;3;1;3;7;3;3;3;2;1;1;0;7;2;3;7;3;3;3;3
;1;3;2;5;6;2;3;1;3;2;3;1;7;3;3;3;7;2;3;3;3;3;3;7;3;3;7;6;7;3;1;5;0;4;3;1;7;4;3;5;7;5;7;2;5;3;4;5
;4;6;6;0;6;6;5;0;6;6;4;6;6;7;6;0;6;0;0;5;0;0;5;5;0;0;0;6;0;0;0;5;5;6;5;5;0;0;6;0;4;0;5;4;0;4;0;4
;5;4;0;4;0;4;0;4;0;0;0;6;5;6;0;4;0;6;6;5;5;5;0;0;4;6].

They are values of features at the third time point of the flow graph.

These results were used to create a temporal flow graph.

Reasoning with Rough Set Flow Graphs

We assume that data sets used to generate rough set flow graphs (Pawlak, 2005) consist of data in tabular forms. From the mathematical point of view, we can treat such entities as information systems (cf. (Pawlak, 1991)). Formally, an information system IS is a quadruple:

$$IS = (U, A, \{V_a\}_{a \in A}, f)$$

where:

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

A special case of an information system is a decision system in which two classes of attributes (the so-called condition attributes and decision attributes) are distinguished, i.e., $A =$

$C \cup D$, where C is the nonempty, finite set of condition attributes, D is the nonempty, finite set of decision attributes. We often consider a decision system with one decision attribute, i.e., $D = \{d\}$.

Any information system can be presented as a data table. Columns of the table are labelled with attributes from the set A, rows are labelled with objects from the set U, and entries of the table are values of the attribute function f assigning to each attribute $a \in A$ and each object $u \in U$, a value of a on u.

In the approach presented in this paper, we consider an information system $IS = (U, A, \{V_a\}_{a \in A}, f\}$ in which a set A of attributes is ordered in time, i.e., $A = \{a_t : t = 1, 2, \ldots, m\}$. The subscript t can be treated as the time point identifier. We assume that the time space is discrete. Such an information system will be called a temporal information system.

The values of attributes can represent different entities, for example, results of measurements, observations or calculations. In the approach presented in this paper, the values of attributes represent clusters of objects, i.e., each value is a cluster identifier.

Figure 2 shows a general idea of acquiring input data. At the beginning, input data consist of objects (in our case, patients) described by sets of features specified in a given time space. Hence, for each object, its temporal characteristic is obtained as it is shown in Figure 2. In general, we can distinguish m time points of interest from the point of view of the problem under consideration. They can be time points when the objects are examined, observed, tested, measured, etc. As it is described later, $m = 3$ in our considered problem.

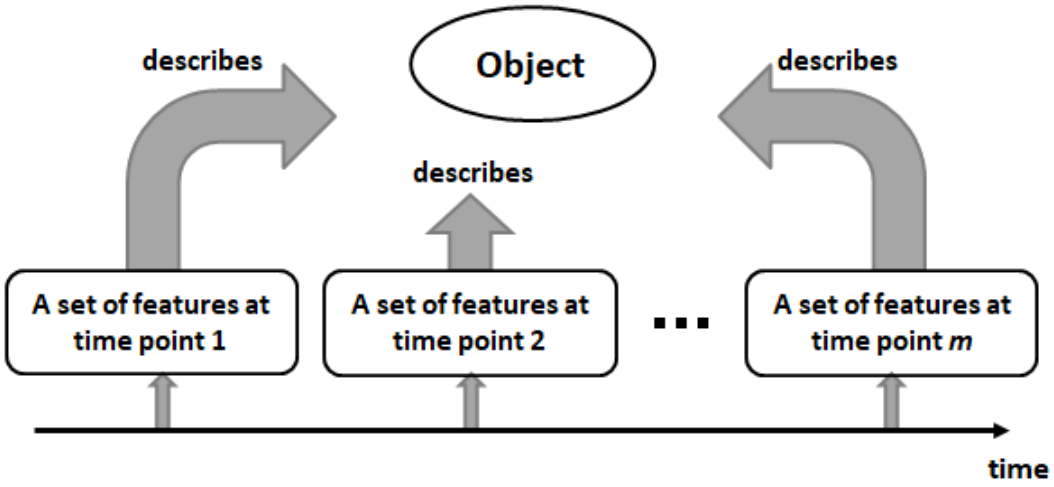

Figure 2. Temporal characteristic of a given object (general).

At each selected time point in a time space, we describe the object by a fixed set of features. It is not necessary for an object to be described by the same sets of features at each time point, i.e., the sets of features can differ both qualitatively and quantitatively. For example, in medical domain, different sets of examinations may be performed or different states may be identified in different situations (e.g., preliminary examinations, check-ups) for a given patient. Sometimes, the check-ups can be more thorough, sometimes less. This can be due to established procedures, capabilities, etc.

Generally, the sets of features considered for objects can be either identical, partly overlapping, or entirely different. The approach presented by us is independent of a particular

situation and it can be used in any of those situations. This is due to the fact that the clustering procedure is made separately over each feature space defined at a given time point. It is worth noting that, at the feature space level, we cannot treat our approach in terms of time series analysis. A time series concerns the same quantities changing in time.

In our approach, we have selected three time points:

1. Before the thermoablation procedure (at this time point, 91 selected features describe patients).
2. 3 months after the thermoablation procedure (at this time point, 29 selected features describe patients).
3. 6 months after the thermoablation procedure (at this time point, 25 selected features describe patients).

Moreover, the fourth time point has been added. It can be identified as evaluation of the cure rate. The attribute corresponding to this time point is a decision attribute (i.e., *d* in a decision system definition). Therefore, we are dealing formally with a decision system as it was defined earlier. A fragment of a temporal information system (including 7 patients) created for the considered data set is shown in Table 1.

Table 1. A fragment of a temporal information system presented in a tabular form.

Attribute a_1	***Attribute a_2***	***Attirbute a_3***	***Decision attribute (d)***
2	4	2	*high*
1	4	2	*medium*
1	7	1	*low*
2	1	3	*medium*
2	4	2	*low*
1	2	3	*medium*
5	7	3	*medium*

Attributes a_1, a_2, and a_3 correspond to the three time points, which were mentioned earlier, respectively. Values of these attributes are cluster identifiers extracted on the basis of proper feature spaces describing patients. Attribute *d* is a decision attribute. For this attribute, a simple interval clustering has been performed and three clusters marked with low, medium and high cure rate, respectively, have been determined. The cuts for determining intervals were as follows: 0.36 and 0.63. These cuts divided the whole range of values of the decision attribute (that is from 0.0 to 1.0) into three intervals corresponding to linguistic labels: low, medium, and high, respectively. For example, the meaning of the first row is as follows. The object (patient), considered in this row, has been assigned by the clustering procedures to the following clusters: 2 (at time point 1), 4 (at time point 2), 2 (at time point 3), and *high* (for decision).

It is worth noting that, at the cluster space level (as, for example, shown in Table 1), we cannot perceive our approach in terms of time series analysis. In this case, we are not dealing with the same quantities varying over time. Each cluster space can be determined over a different feature space. Therefore, the domains (the sets of values) of attributes in a temporal information system cannot be equated with each other.

One can see that, in general, we can assign a linguistic description to each cluster identifier, which is important, for example, from the point of view of medical diagnosis.

Any sequence $< v_1, v_2, \ldots, v_k >$, where $v_1 \in V_{a_j}$, $v_2 \in V_{a_{j+1}}$, ..., $v_k \in V_{a_{j+k-1}}$, $j = 1,2, \ldots, m$, and $j + k - 1 \leq m$, is called an episode (over the sets of values of attributes) in a temporal information system $IS = (U, A, \{V_a\}_{a \in A}, f\}$. Two episodes $< v_1, v_2, \ldots, v_k >$ and $< v'_1, v'_2, \ldots, v'_k >$ are adjacent if and only if either $v_1 = v'_k$ or $v_k = v'_1$. In our approach, episodes are defined over sets of clusters. Basically, one episode is a sequence of cluster identifiers (see Figure 3).

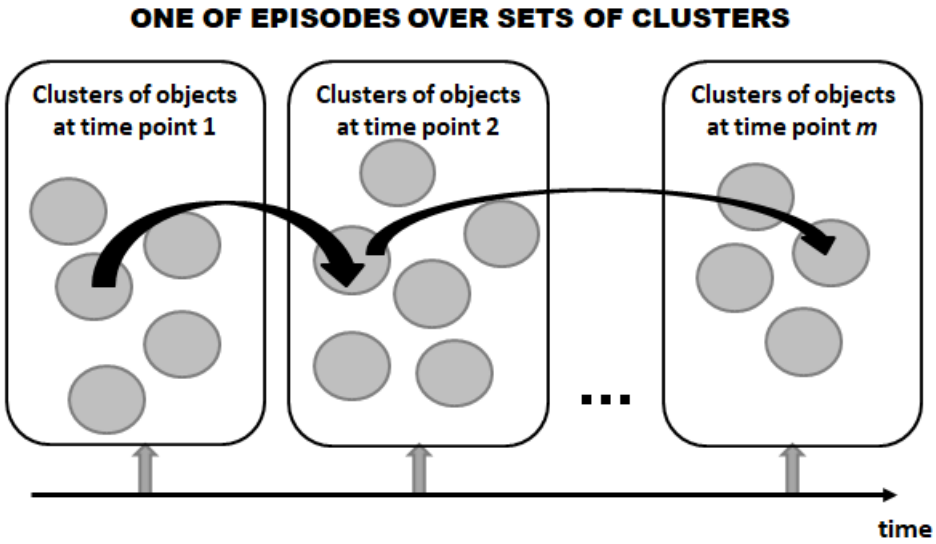

Figure 3. An episode defined over sets of clusters.

In Figure 4, a general scheme of the proposed procedure to build a flow graph and extract episodes is shown. As a result of the procedure, we obtain the episodes, which can be used to make temporal inference about characteristics of objects. In the first step, we determine sets of clusters of objects. The clusters are extracted with respect to feature spaces defined at selected time points. In this step, the ant-based clustering algorithm can be used as it was described in the previous section. As a result of this step, we obtain an information about cluster identifiers to which objects have been assigned in each feature space selected for a given time point. This information is arranged as a temporal information system defined earlier. On the basis of this system, the so-called rough set flow graph is built. Figure 5 shows the proposed procedure applied to data taken into consideration by us.

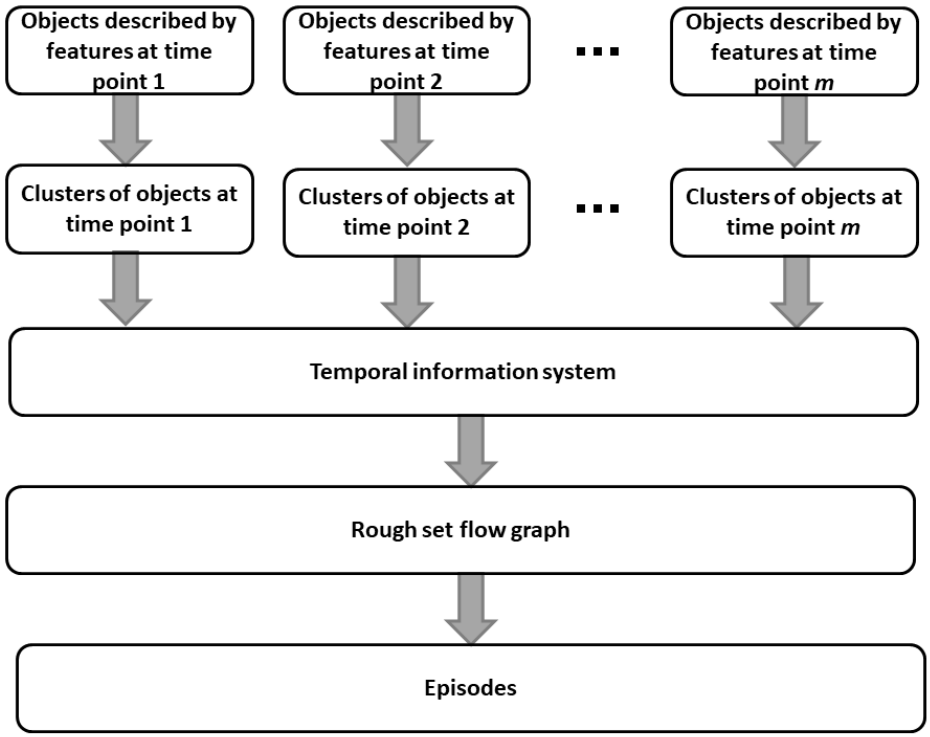

Figure 4. A general scheme of the proposed procedure to build a flow graph.

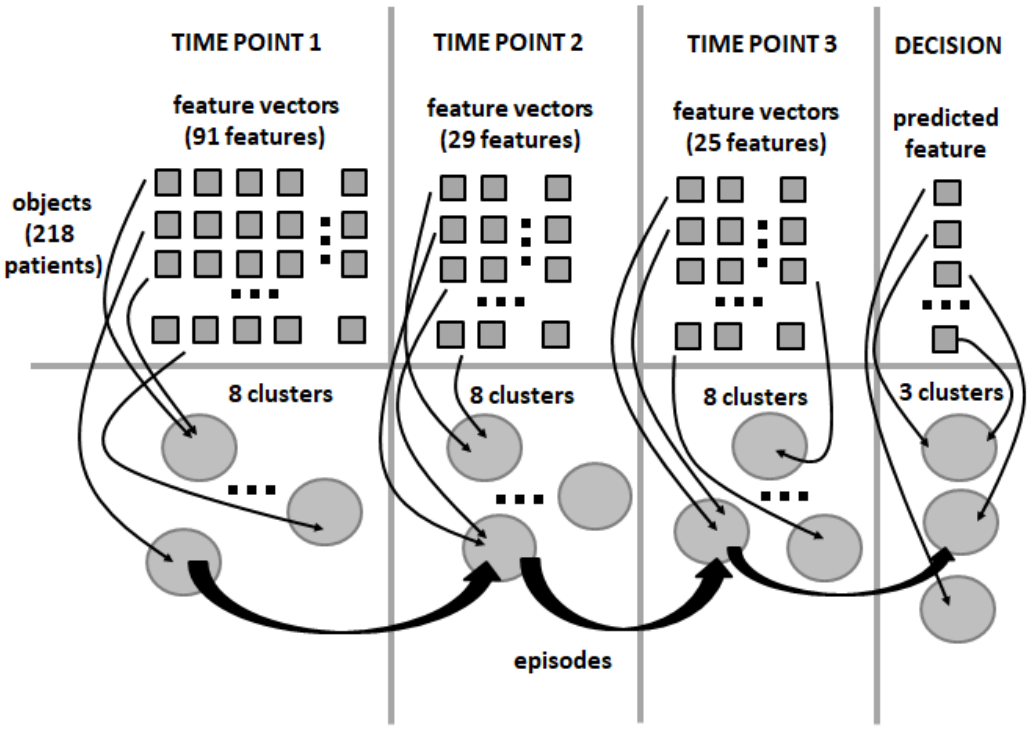

Figure 5. The procedure applied to data taken into consideration.

Information flow distribution is the kind of knowledge that can be helpful in solving different problems appearing in data analysis, especially, if we deal with temporal data, i.e., results of observations or measurements are ordered in time, (cf. (Pancerz, Lewicki, Tadeusiewicz, & Warchol, 2013)). In the literature, different approaches based on flow graphs were proposed. The fundamental one, called flow networks, was proposed by L.R. Ford and D.R. Fulkerson (Ford & Fulkerson, 1962).

Rough set flow graphs were defined by Z. Pawlak (Pawlak, 2005) as a tool for reasoning from data. Let $IS = (U, A, \{V_a\}_{a \in A}, f\}$ be a temporal information system with $A = \{a_t : t = 1,2, \dots, m\}$. A rough set flow graph corresponding to *IS* is a tuple:

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

where:

- $N = N_{a_1} \cup N_{a_2} \cup \dots \cup N_{a_t}$ is the set of nodes such that for each $a \in A$: $card(N_a) = card(V_a)$, i.e., each node in N_a corresponds exactly to one attribute value from V_a,
- $B \subseteq N \times N$ is the set of multi-labelled directed branches, where each branch $(n, n') \in B$, where $n \in N_{a_{i-1}}$, $n' \in N_{a_i}$, and $i = 2,3, \dots, t$, is labelled with three values:
  - $cer(n, n') = \frac{card(\{u \in U : f(a_{i-1}, u) = v \wedge f(a_i, u) = v'\})}{card(\{u \in U : f(a_{i-1}, u) = v\})}$,
  - $cov(n, n') = \frac{card(\{u \in U : f(a_{i-1}, u) = v \wedge f(a_i, u) = v'\})}{card(\{u \in U : f(a_i, u) = v'\})}$,
  - $str(n, n') = \frac{card(\{u \in U : f(a_{i-1}, u) = v \wedge f(a_i, u) = v'\})}{card(U)}$,

  such that *n* corresponds to $v \in V_{a_{i-1}}$ and *n'* corresponds to $v' \in V_{a_i}$.

These three values $cer(n, n')$, $cov(n, n')$, and $str(n, n')$ are called certainty, coverage, and strength of the branch (n, n'), respectively. One can see that $cer(n, n') \in [0,1]$, $cov(n, n') \in [0,1]$, and $str(n, n') \in [0,1]$ for each *n* and *n'*. These coefficients assigned to each branch play an imporant role in reasoning from data. One can see that formulas determining certainty and covering refer to Bayes' rules. Hence, rough set flow graphs can be used as an interesting tool to perform resoning processes.

The algorithm for creating a temporal rough set flow graph corresponding to a temporal information system can be described as follows:

1. At the beginning, an empty set *N* of nodes and an empty set *B* of directed branches are created.
2. For each attribute *a* in the set *A* of attributes of a temporal information system, for each value *v* of *a*, the node corresponding to *v* is created and added to *N*.
3. For each pair of nodes corresponding to values of adjacent attributes in a temporal information system, a branch linking these nodes is created and added to *B*.
4. For each branch, three labels associated with it are calculated on the basis of formulas given earlier (see definitions of certainty, coverage, and strength).

It is worth noting that, in case of a rough set flow graph built on the basis of a temporal information system, each layer of nodes in the graph corresponds to values of attributes determined for a given time point.

The rough set flow graph obtained for the considered data set is shown in Figure 6 (this graph is extracted from a fragment of a temporal information system shown in Table 1). Each

horizontal layer includes nodes corresponding to clusters obtained for a given time point. Each node is marked with the attribute name and the cluster identifier. Each branch is described by three coefficients: certainty (*cer*), coverage (*cov*), and strength (*str*) mentioned earlier.

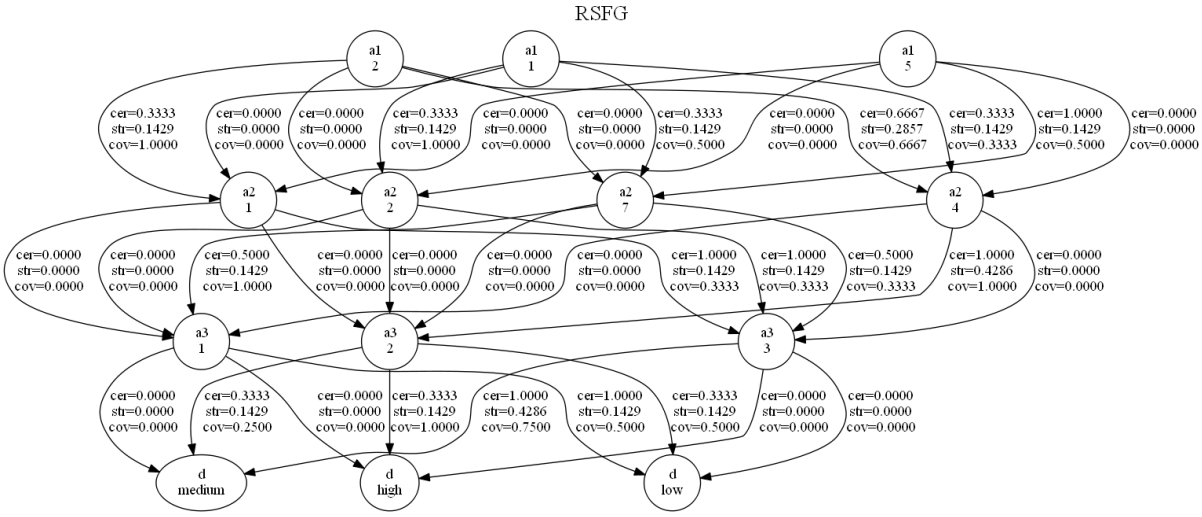

Figure 6. A fragment of the rough set flow graph.

The final goal of the presented approach is to find significant episodes over the sets of clusters. Discovering such episodes can be made by the approach using the ant-based clustering algorithm as it was shown in (Lewicki & Pancerz 2020). The main goal of the ants is to carry episodes and concatenate them into longer episodes. We can say that the longer the episode, the more information it gives to us. The longer episode represents the longer path of possible states of objects in consecutive time points.

The ant-based algorithm can be described as follows.

1. At the beginning, the episodes, each consisting of one cluster, are created.
2. A given ant picks up randomly one of the episodes and carries it.
3. An episode can be dropped by the ant next to another episode if these episodes are adjacent. Then a new episode is created by concatenating these two episodes together. The probability of concatenation depends on the certainty of the new episode.

It is worth noting that to determine the certainty of the new episode created during the clustering procedure, we can use a special class of functions, called triangular norms (known especially from fuzzy set theory) due to the fact that $cer(n, n') \in [0,1]$ for any nodes *n* and *n'*. Triangular norms can be either t-norms or t-conorms (see (Klement, Mesiar & Pap, 2000). In general, the certainty of the new episode *e* is a triangular norm *T* of the certainties of the concatenated episodes *e'* and *e''*, i.e.:

$$cer(e) = T(cer(e'), cer(e'')).$$

Moreover, we can use strength or coverage coefficients instead of a certainty coefficient.

The rough set flow graph from Figure 6 with a distinguished episode is shown in Figure 7.

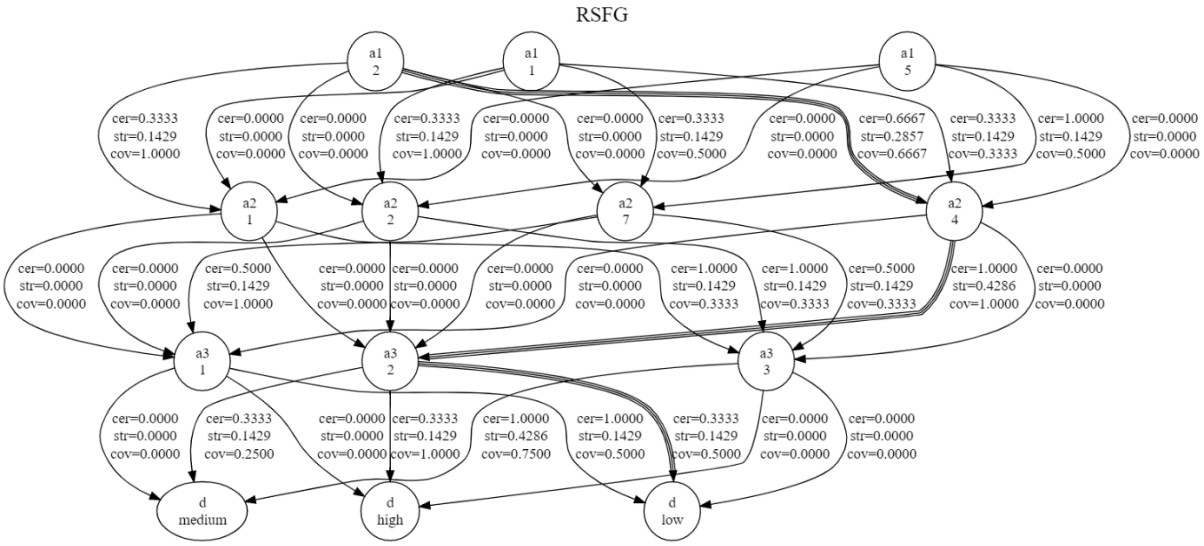

Figure 7. A fragment of the rough set flow graph with a distinguished episode.

The fragment of the set of episodes extracted by means of the ant-based algorithm (in which a minimum t-norm was used to determine certainties of created episodes during the clustering procedure) is as follows:

a2=3 -> a3=6 -> d=high
a3=4 -> d=high
a2=7 -> a3=7 -> d=medium
a1=1 -> a2=1 -> a3=3 -> d=medium
a3=5 -> d=high
a3=0 -> d=high
a1=4 -> a2=2 -> a3=2 -> d=low
a2=2 -> a3=1 -> d=medium
a1=2 -> a2=1 -> a3=3 -> d=medium
a1=0 -> a2=0 -> a3=7 -> d=medium
a1=4 -> a2=2 -> a3=2 -> d=medium
a2=4 -> a3=2 -> d=medium
a1=7 -> a2=6 -> a3=3 -> d=medium
a1=4 -> a2=2 -> a3=3 -> d=medium
a1=5 -> a2=5 -> a3=3 -> d=medium

Let us consider episode:

a2=4 -> a3=2 -> d=medium

In the considered data set, this episode is supported by the patients included in Table 2. Having this information, we can move to description of clusters: 4 (at time point 2) and 2 (at time point 3) and analyse values of features taken into consideration in these time points.

Table 2. A fragment of a temporal information system including patients supporting episode a2=4 -> a3=2 -> d=medium

Attribute a_1	***Attribute a_2***	***Attirbute a_3***	***Decision attribute (d)***
1	4	2	medium
1	4	2	medium
5	4	2	medium
2	4	2	medium
2	4	2	medium

The effectiveness of the algorithm was validated using the ten-fold cross validation approach. In each iteration, nine parts were used in the training stage to create a rough set flow graph and to generate significant episodes using the ant-based clustering. One part was used to determine actual certainties of the episodes found in the training stage. Next, we calculated: the average value of actual certainties of the episodes, the average value and the standard deviation for differences between the actual certainties of the episodes and the predicted certainties (determined in the training step) of the episodes.

The presented algorithm has the polynomial time complexity. One can see that the brute force search of the entire space of episodes leads to the exponential time complexity on account of a number of attributes in a given temporal information system.

The extracted episodes can be used in reasoning processes. Each episode can be seen as rules delivering information flow between particular layers corresponding to time points in a given time space. In contrast to the classical classification problem, the presented approach makes it possible to gain information about the possible successions of states between stages defined in a given time space. Such successions of states are expressed by episodes. Therefore, not only reasoning on final decisions is possible. The local temporal inference is also available.

Software Tool

Creation of rough set flow graphs and extraction of episodes using the ant-based clustering algorithm have been implemented in our software tool called Classification and Prediction Software System (CLAPSS), see (Pancerz, Lewicki & Sarzynski, 2019). It is a tool developed for solving different classification and prediction problems using, among others, some specialized approaches based mainly on fuzzy sets and rough sets. The tool is equipped with a graphical interface (see Figure 8). The ant-based clustering procedure for extracting episodes can be parametrized by the user as it is shown in Figure 9. The user can set the number of ants and the number of iterations. Moreover, one can select one of the triangular norms from the sets of the most known t-norms and t-conorms (minimum/maximum, algebraic/probabilistic product, Lukasiewicz, Einstein, Hamacher, Fodor, drastic).

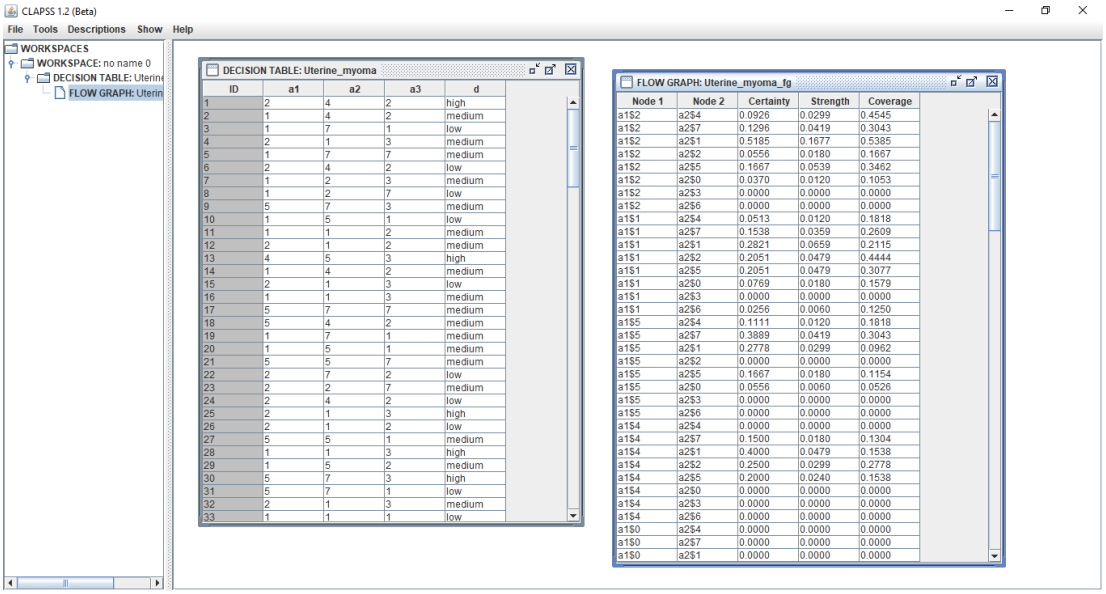

Figure 8. The GUI of CLAPSS.

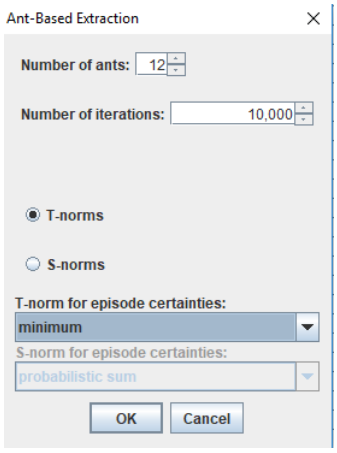

Figure 9. Parametrization of the ant-based clustering procedure for extracting episodes.

An information or decision system (particularly, a temporal information system) can be loaded into CLAPSS from files recorded in several formats. The tool accepts popular text formats of data tables used in other data mining and machine learning tools, i.e., WEKA (Hall et al., 2009), RSES (Bazan & Szczuka, 2005) and the XML format used in ROSETTA (Ohrn, Komorowski, Skowron & Synak, 1998). CLAPSS enables the user to export the created rough set flow graph to the DOT format used in the Graphviz tool (Ellson, Gansner, Koutsoos, North & Woodhull, 2004) as well as to the PNG format. Inputs and outputs in CLAPSS are shown in Figure 9. One can see interoperability of CLAPSS with other software tools.

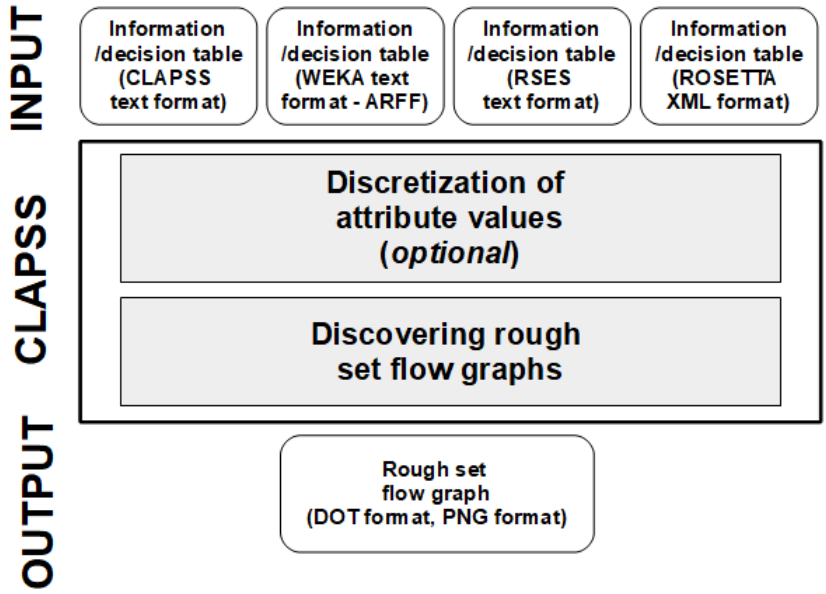

Figure 10. Inputs and outputs in CLAPSS.

The fragment of a temporal information system presented in Table 1, recorded in the internal CLAPSS format, looks like this:

TABLE Uterine_myoma
a1 a2 a3 d
integer integer integer string
condition condition condition condition
2 4 2 high
1 4 2 medium
1 7 1 low
2 1 3 medium
2 4 2 low
1 2 3 medium
5 7 3 medium

DISCUSSION AND CONCLUSIONS

Many classical medical decision problems are reduced to classification problems based on sets of feature vectors describing patients. In the paper, we have shown the approach based on temporal information about patients. Results of diagnostics during the treatment process (multi-stage in many cases) can be used to model temporal information flow and to perform temporal inference. It is possible due to separation of features determined at different stages of treatment.

Moreover, many classical medical decision problems are focused on the construction of the most accurate classifiers or classifier ensembles. In this case, the main goal is to obtain the

best quality of the classification of patients. Accurate classification is a significant problem, but other important issues considered in medical diagnosis are: explanation of decisions made for new patients and imitation of human diagnostic procedures by computer tools. Then, solving a classification problem is not the only goal. The diagnosticians also strive to understand, among other things, the structure of the decision problem, interdependencies between particular stages of treatment, etc. As it has been presented in the paper, extracted episodes have a human readable form that can be used in reasoning processes. Our software tool (CLAPSS) enables the users to analyse medical decision problems. The models in the form of rough set flow graphs, as well as extracted episodes, enrich the problem analysis, and ultimately the diagnosis.

In our approach, we have used the ant-based clustering algorithms which are capable of dealing with the data with a poorly known structure. It is important in case of dozens of features describing patients. As the example, the problem of the thermoablation procedure in treatment of uterine myoma disease has been considered. However, in the paper, we have shown a general scheme for the modelling of the decision problems in case of multistage treatments. The presented approach fits into the general problems considered in data mining called sequence mining or temporal data mining (cf. Dong & Pei, 2007; Mitsa, 2010).

The presented approach differs from the approaches to discovering frequent episodes. We can indicate the following significant differences:

1. The frequent episodes are considered as collections of events that occur relatively close to each other in a given partial order (cf. (Mannila, Toivonen & Verkamo, 1997)). Particularly, serial and parallel episodes are considered. In our approach, we consider only ordered sets of attribute values describing objects of interest.
2. Our approach is focused on certainties of sequences, not on their frequencies of occurrence. In our approach, we are interested in information flow between elements of episodes (coefficients associated with branches, linking nodes in rough set flow graphs). If the frequencies of occurrence are considered, the coexistence of elements in episodes is taken into account.

The presented approach differs also from the approaches to clustering episodes. Classical clustering procedures implement the so-called vertical clustering (episodes are clustered to build groups of episodes). In our approach, a clustering procedure implements the so-called horizontal clustering. Shorter pieces of episodes (starting with one-element episodes) are concatenated to build longer ones (according to given criteria).

One of our directions for further research is to apply fuzzy flow graphs to the considered and other medical decision problems. Creation of fuzzy flow graphs and extraction of episodes using the ant-based clustering algorithm have been also implemented in our software tool (CLAPSS), see (Pancerz, Lewicki & Sarzynski, 2019). Moreover, an important issue is to use a domain knowledge to improve solving decision problems (see for example (Bazan, Buregwa-Czuma & Jankowski, 2013)). The domain knowledge can have different forms. One of them has the form of ontologies. Linking ontologies with attribute values in temporal information systems was considered in (Pancerz, 2016).

REFERENCES

Alnoukari, M., & El Sheikh, A. (2012). Knowledge discovery process models: from traditional to agile modeling. In *Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications* (pp. 72-100). IGI Global.

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

Bazan, J. G., & Szczuka, M. (2005). The rough set exploration system. In *Transactions on Rough Sets III* (pp. 37-56). Springer, Berlin, Heidelberg.

Boryzczka, U. (2008). Ant clustering algorithm: Intelligent information systems. Kluwer Academic Publishers.

Burney, A., & Abbas, Z. (2015). Applications of rough sets in health sciences and disease diagnosis. *Recent Researches in Applied Computer Science*, *8*(3), 153-161.

Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V., & Zou, J. (2021). Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review. *JAMA dermatology*, *157*(11), 1362-1369.

De La Cruz, M. S. D., & Buchanan, E. M. (2017). Uterine fibroids: diagnosis and treatment. *American family physician*, *95*(2), 100-107.

Deneubourg, J. L., Goss, S., Franks, N., Sendova-Franks, A., Detrain, C., & Chrétien, L. (1991, February). The dynamics of collective sorting robot-like ants and ant-like robots. In *From animals to animats: proceedings of the first international conference on simulation of adaptive behavior* (pp. 356-365).

Dhillon, A., & Singh, A. (2019). Machine learning in healthcare data analysis: a survey. *Journal of Biology and Today's World*, *8*(6), 1-10.

Dong, G., & Pei, J. (2007). *Sequence data mining* (Vol. 33). Springer Science & Business Media.

Ellson, J., Gansner, E. R., Koutsofios, E., North, S. C., & Woodhull, G. (2004). Graphviz and dynagraph—static and dynamic graph drawing tools. In *Graph drawing software* (pp. 127-148). Springer, Berlin, Heidelberg.

Ford, L. R., & Fulkerson, D. R. (2015). *Flows in networks*. Princeton university press.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.

Handl, J., Knowles, J. D., & Dorigo, M. (2003). On the Performance of Ant-based Clustering. *HIS*, *105*, 204-213.

Kalamarz, P., Zagrobelna, M., & Pyziak, L. (2017). Focusing ultrasounds beam. *Physics for Economy*, *1*, 15-26.

Klement, E. P., Mesiar, R., & Pap, E. (2000). Families of t-norms. In *Triangular Norms* (pp. 101-119). Springer, Dordrecht.

Lewicki, A., & Pancerz, K. (2020). Ant-based clustering for flow graph mining. *International Journal of Applied Mathematics and Computer Science*, *30*(3), 561-572.

Lozinski, T., Filipowska, J., Pyka, M., Baczkowska, M., & Ciebiera, M. (2021). Magnetic resonance-guided high-intensity ultrasound (MR-HIFU) in the treatment of symptomatic uterine fibroids—five-year experience. *Ginekologia Polska*. doi: 10.5603/GP.a2021.0098.

Lumer, E. D., & Faieta, B. (1994, July). Diversity and adaptation in populations of clustering ants. In *Proceedings of the third international conference on Simulation of adaptive behavior: from animals to animats 3: from animals to animats 3* (pp. 501-508).

Mannila, H., Toivonen, H., & Verkamo, A. I. (1997). Discovery of frequent episodes in event sequences. *Data mining and knowledge discovery*, *1*(3), 259-289.

Mitsa, T. (2010). *Temporal data mining*. CRC Press.

Øhrn, A., Komorowski, J., Skowron, A., & Synak, P. (1998). The ROSETTA software system. *Rough Sets in Knowledge Discovery*, *2*, 572-576.

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

Pancerz, K., Lewicki, A., & Tadeusiewicz, R. (2012, July). Ant based clustering of two-class sets with well categorized objects. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 241-250). Springer, Berlin, Heidelberg.

Pancerz, K., Lewicki, A., Tadeusiewicz, R., & 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., Lewicki, A., & Sarzyński, J. (2019, June). Discovering Flow Graphs from Data Tables Using the Classification and Prediction Software System (CLAPSS). In *International Joint Conference on Rough Sets* (pp. 356-368). Springer, Cham.

Pawlak, Z. (1991). *Rough sets: Theoretical aspects of reasoning about data* (Vol. 9). Springer Science & Business Media.

Pawlak, Z. (2005). Flow graphs and data mining. In *Transactions on rough sets III* (pp. 1-36). Springer, Berlin, Heidelberg.

Pawlak, Z. (2005, August). Rough sets and flow graphs. In *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing* (pp. 1-11). Springer, Berlin, Heidelberg.

Pawlowski, C. (2019). *Machine learning for problems with missing and uncertain data with applications to personalized medicine* (Doctoral dissertation, Massachusetts Institute of Technology).

Ren, S., Lu, X., & Wang, T. (2018, March). Application of ontology in medical heterogeneous data integration. In *2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)* (pp. 150-155). IEEE.

Sainio, T., Saunavaara, J., Komar, G., Mattila, S., Otonkoski, S., Joronen, K., ... & Blanco Sequeiros, R. (2021). Feasibility of apparent diffusion coefficient in predicting the technical outcome of MR-guided high-intensity focused ultrasound treatment of uterine fibroids–a comparison with the Funaki classification. *International Journal of Hyperthermia*, *38*(1), 85-94.

Salcedo-Bernal, A., Villamil-Giraldo, M. P., & Moreno-Barbosa, A. D. (2016). Clinical data analysis: An opportunity to compare machine learning methods. *Procedia Computer Science*, *100*, 731-738.

Singh, S., & Srivastava, S. (2020). Review of Clustering Techniques in Control System: Review of Clustering Techniques in Control System. *Procedia Computer Science*, *173*, 272-280.

Su, T., & Dy, J. G. (2007). In search of deterministic methods for initializing K-means and Gaussian mixture clustering. *Intelligent Data Analysis*, *11*(4), 319-338.

Thirumahal, R., & Sadasivam, G. (2020). Data integration techniques for healthcare – a comprehensive survey. *International Journal of Computer Sciences and Engineering Open Access*.

Tripathi, S., Muhr, D., Brunner, M., Jodlbauer, H., Dehmer, M., & Emmert-Streib, F. (2021). Ensuring the robustness and reliability of data-driven knowledge discovery models in production and manufacturing. *Frontiers in Artificial Intelligence*, *4*, 22.

Verpalen, I. M., Anneveldt, K. J., Nijholt, I. M., Schutte, J. M., Dijkstra, J. R., Franx, A., ... & Boomsma, M. F. (2019). Magnetic resonance-high intensity focused ultrasound (MR-HIFU) therapy of symptomatic uterine fibroids with unrestrictive treatment protocols: A systematic review and meta-analysis. *European journal of radiology*, *120*, 108700. doi: 10.1016/j.ejrad.2019.108700.

Vlamou, E., & Papadopoulos, B. (2019). Fuzzy logic systems and medical applications. *AIMS neuroscience*, *6*(4), 266.

Wen, D., Khan, S. M., Xu, A. J., Ibrahim, H., Smith, L., Caballero, J., ... & Matin, R. N. (2021). Characteristics of publicly available skin cancer image datasets: a systematic review. *The Lancet Digital Health*.

Authors' Note

All correspondence should be addressed to
Arkadiusz Lewicki
Faculty of Applied Computer Science
University of Information Technology and Management
Sucharskiego Str. 2, 35-225 Rzeszow, Poland,
alewicki@wsiz.edu.pl

Human Technology
ISSN 1795-6889
https://ht.csr-pub.eu