ARES SYSTEM - INTEGRATION OF ANALYSIS METHODOLOGIES

Roman Podraza, Mariusz Kalinowski

Abstract: ARES System is an application dedicated to data analysis, data exploration and knowledge discovery. This versatile tool offers rough set based methodology as well as emerging patterns and Support Vector Machine algorithms to be applied to input data presented as an information system. Many analytical capabilities provided with an user-friendly, intuitive graphical interface and XML based support for input/output operations make ARES System a challenging and promising software for application in scientific research and didactic work. An unique feature of ARES System is its ability of identifying improper objects within information system by employing credibility coefficients. The credibility coefficients attempt to assess a degree of typicality of each object in respect to the whole information system.

Keywords: information system, classification, rough sets, emerging patterns, SVM, credibility coefficients.

ACM Classification Keywords: H.1.1 Systems and Information Theory

Introduction

ARES System has been persistently developed to extend its potential in data analysis, data exploration and knowledge discovery [Podraza, 2005]. Its initial functionalities were based on rough set theory [Pawlak, 1991] and contained all phases leading to discovering rules from an information system. Results of each step of data analysis can be presented by an intuitive graphical user interface. This feature should enable user to observe the whole process of data analysis and/or exploration to learn how to tune the whole experiment to achieve the most sound results. A number of algorithms of data discretization, finding frequent sets and reducts or extracting rules were implemented with the first release of the system.

From the very beginning ARES System offered credibility coefficients [Podraza, 2007], [Podraza 2006] , which were heuristic measures of objects' typicality in respect to other ones. The main idea of introducing the credibility coefficients could be summarized that rules involve exceptions to them and the exceptions very often can be more interesting that the rules themselves. The evaluation of credibility coefficients is based on observations that typical objects are considered credible, because they appear more frequently and we used to generalize their appearance as something predictable, and describable by rules.

In the development of ARES System there have been introduced algorithms for discovering discriminants of information systems – LEM1, LEM2 and AQ along with required checking of consistency of information systems [Grzymala, 2005]. A next new feature of ARES System enables discovering Emerging Patterns (EP) in objects of decision tables. This is an alternative approach to mine rules from set of data. Support Vector Machine is yet another methodology used to classify data, which has been integrated within the system.

ARES System becomes a common platform for a number of different classification approaches. The same data can be processed in several ways and the results can be immediately compared and examined. Each approach can be parameterized in many aspects giving a really flexible and powerful classification tool. Formats of input/output files have been enriched by flexible XML based enabling universal communication with other different systems. In particular quite large system output like hundreds of rules can be presented in web browser or more elaborated presentation tools.

In the next section a general description of ARES System functionalities can be found. Then follows a section covering capabilities of three methodologies of data analysis and knowledge acquisition, which have been introduced to ARES System. A brief presentation of system development goes after a short record of the original

system potential. Subsequently credibility coefficients, which are unique feature o ARES System, are concerned. And finally conclusions and further perspectives of ARES System are submitted.

System Overview

The system has been designed to give a full interactive access to process of data analyses involving different approaches. A multi-window graphical user interface enables tracking different steps of the processing and/or comparing results of applying different algorithms for the whole procedure of data analysis or for its particular phase. A sample screenshot from ARES System is shown in Fig. 1. The main application window consists of two views. *Directory browser* presents a list of all elements currently available for user (e.g. workspaces, decision tables, analysis results) and *Workspace view* provides a space for windows opened according to user selection. Views of elements from the directory browser can be presented for investigations and comparisons with the others (e.g. window with a decision table, the corresponding frequent sets, a list of all found reducts).

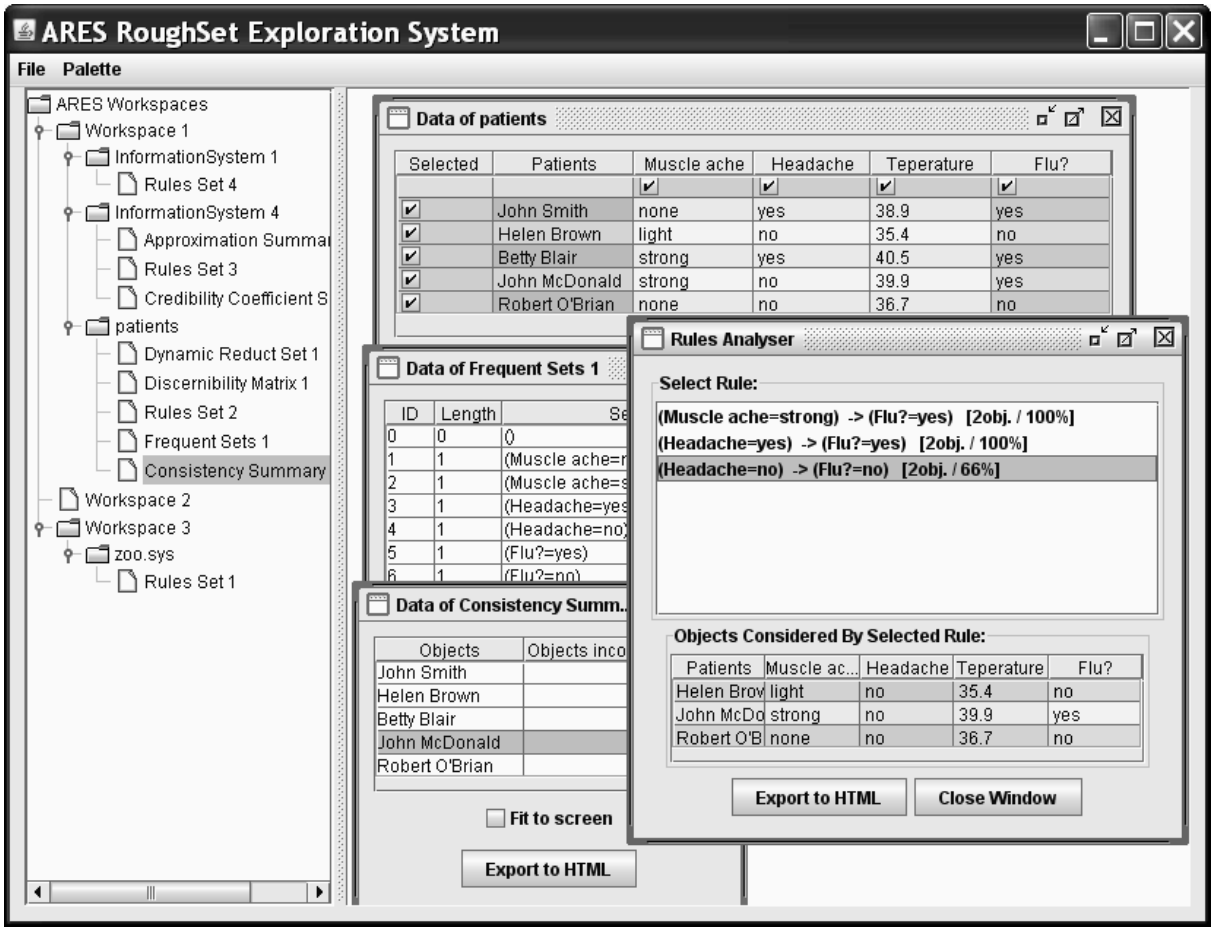


Figure 1. Multi-window GUI of ARES System

A relationship between elements of processed data is reflected by their hierarchical placements within the system directory, however it can be accentuated by naming conventions. Each directory item stores information on its type, its data and description. Item type determines operations, which can be applied to it. Context menu associated with the item (pulled down by the right button of mouse) contains positions for starting the operations. The data and the description can be presented in appropriate windows in the workspace view. The description contains some statistic and explanatory information as execution time used to produce the item, algorithm applied and some other specific data (e.g. number of generated rules for rule set item).

An information system or more precisely a decision table is input to ARES System. The input data can be stored in three formats: numerical one (specific for ARES System only), CSV and XML. Two latter ones give sufficient

flexibility to cooperate with other systems. In the decision table rows represent data objects and columns represent attributes of objects. In ARES System only one attribute can be chosen as a decision. There are checkboxes associated with each row and each columns. They enable cutting a desired decision table from the existing one by removing unchecked rows and/or columns (Fig. 2). It is possible to view the information system in its internal numerical representation as well.

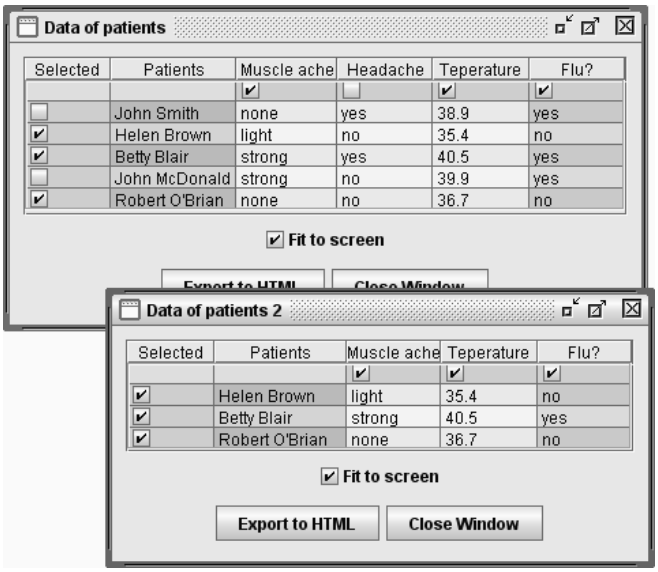


Figure 2. Modifications of a Decision Table

All data displayed in a tabular form in any window of the workspace view can be exported to HTML file. The information format is similar to presented in the window. This option allows saving the results of ARES System, present them and print in internet browsers and/or publish in formats accepting HTML.

Integration of Platforms

The system has been developed to comprise three methodologies of knowledge acquisition. To original contents based on Rough Set theory [Pawlak, 1991] capabilities of Emerging Patterns [Dong, 1999] approach and Support Vector Machines [Schmilovici, 2005] methodology have been appended to the system. Rough Set approach has been extended as well by adding algorithms for discovering discriminant of information system. SVM algorithms have already been implemented and user interface for them is just under construction, and currently only credibility coefficients based on SVM are fully operational. The original ARES System [Podraza 2005] has comprised modules for performing the following tasks form rough set sphere such as

- Discretizing continuous domains of objects' attributes
- Discovering approximations of decision classes
- Determining discernibility matrices
- Finding relative reducts by applying set of algorithms to calculate
  - all reducts
  - minimal reducts
- Discovering frequents sets
- Mining decision rules

All these tasks can be performed by a number of algorithms. The operations are selected for particular items from the directory browser. Each such item has a number of operations applicable to it. They are all grouped in a context menu pulled down by the right button of mouse while selecting the item. A collection of operations applicable to directory item representing a set of rules is presented in Tab. 1. Usually results of each such

operation is presented in a new window in the workspace area and sometimes a new item representing the result is inserted to the directory. There is a possibility to export the result into HTML file. Particular operation can be performed by a number of algorithms. The choice and required parameters are determined in a interactive way.

Table 1. List of commands (from context menu) for item representing set of rules

Command	Description
Close Rules Miner	Deletes the rules set
Show Data	Displays a window with rules and information about them (support and confidence of given rules)
Show Properties	Displays a window with information about given rules set (name and parameters of the rule mining algorithm, count of rules, time of mining)
Show Disjunctive Rules Set (AQ)	Displays a window with disjunctive rules (generated by the AQ algorithm)
Objects Coverage	Displays a window presenting objects coverage
Rules Coverage	Displays a window presenting rules coverage
Rules Analyser	Displays a window of a rules analyser
Show All Rules	Displays a window with all rules
Show Certain Rules	Displays a window with certain rules
Show Possible Rules	Displays a window with possible rules
Close Menu	Closes the menu

In implementation of ARES System many known procedures of rough set area have been instantiated and some updates have been formulated. There are some analyzing tools to present relationships discovered between results and objects of original decision table. In Fig. 3 there is a window showing objects from decision table supporting antecedent of a selected rule.

There have been a number of algorithms for calculating credibility coefficients for objects of decision table. Module of credibility coefficients is very special for ARES System because it allows for systematic treatment of exceptional cases and the next section is devoted to this subject in a systematic way.

A domain of rough set theory in ARES System has been supplemented by module for discovering discriminant of information system. The module comprises three algorithms LEM1, LEM2 and AQ [Grzymala, 2005]. The last algorithm generates rules with disjunctive representation. There is a possibility to check consistency of information system, which is a necessary condition to calculate a discriminant of information system. If the information system is inconsistent (there are at least two objects, which have the same values for all conditional attributes and have different values of the decision attribute) then the operation results in a report highlighting objects causing this inconsistency.

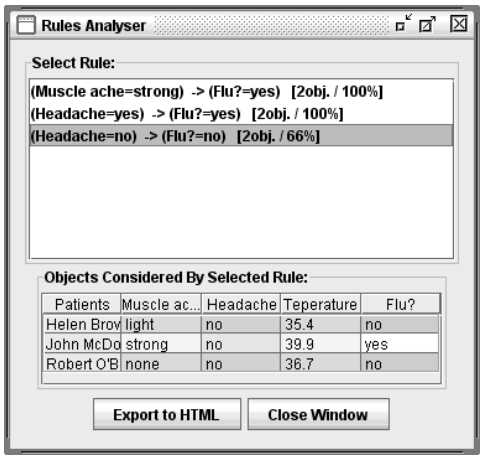


Figure 3. Window presenting relationship between antecedent of selected rule and objects from decision table.

The next platform in ARES System is the KTDA system [Podraza 2007], [Podraza, 2006a] based on Emerging Patterns (EP) approach. The KTDA system was designed and implemented as an independent platform and then was integrated with ARES System. To distinguish two decision classes it is desirable to find out such patterns which are frequent itemsets in one class and are infrequent in the other one. These patterns are just called emerging patterns. The ratio of the pattern support in its target class to the pattern support in the rest of the dataset is a *growth rate* for this pattern. Larger values of the growth rate denote more characteristic EPs for its target class.

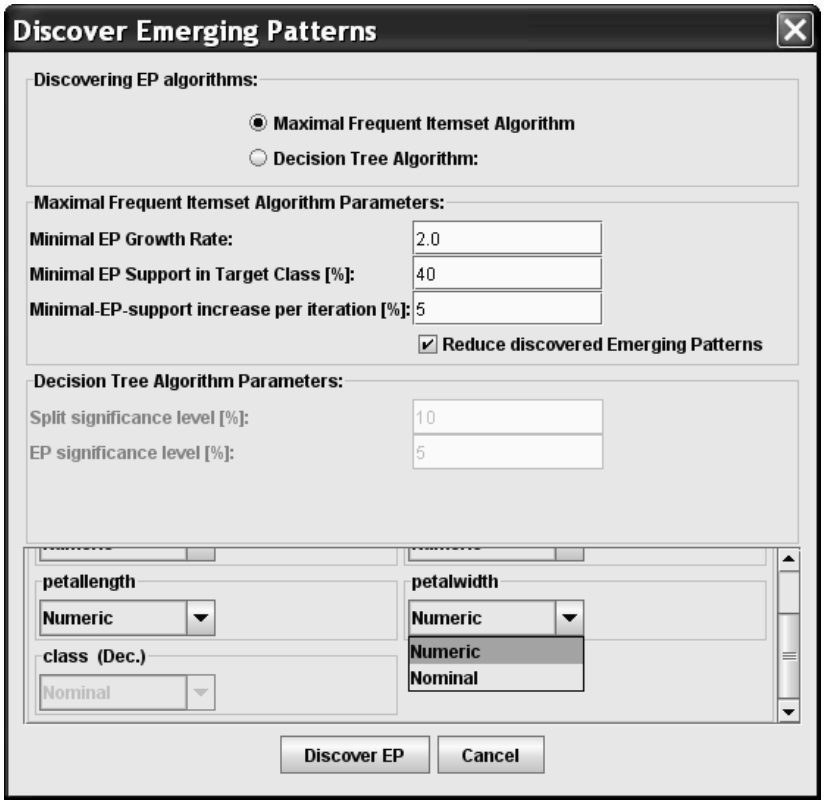
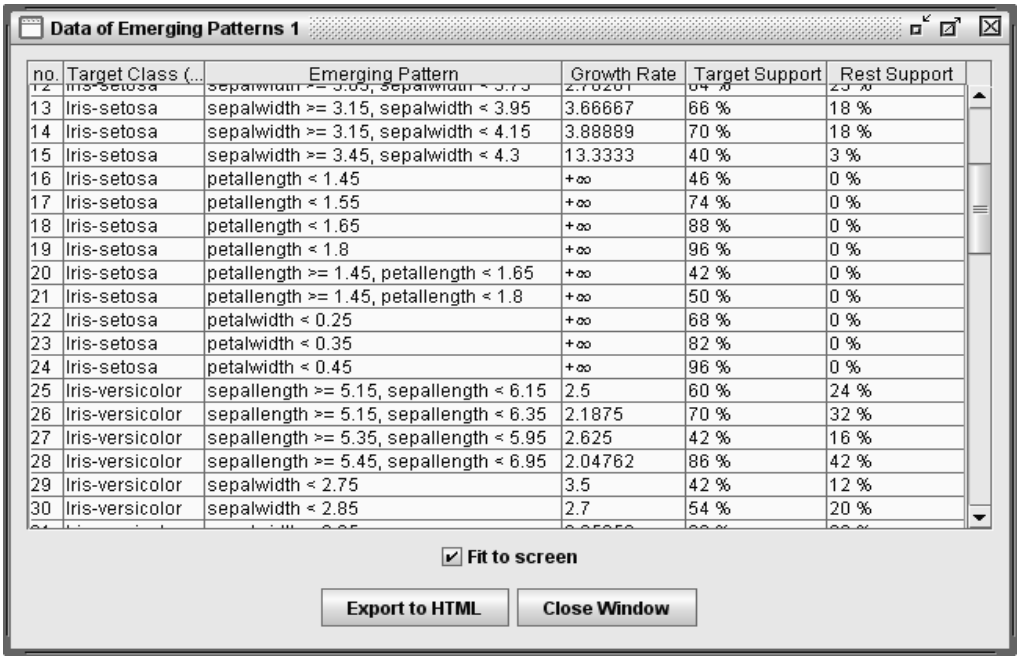


Figure 4. A choice of algorithms for discovering emerging patterns.

The KTDA system implements two different algorithms of discovering EPs – using maximal frequent itemsets proposed in [Dong, 1999] and using decision tree [Boulesteix, 2003] , but with some extensions and improvements. The former one reflects the classical approach and requires stating minimal growth rate and minimal support in the target class, while the latter one uses Fisher's Exact Test used to discover only such EPs which are statistically significant. Algorithm using decision tree is quicker one, produces smaller set of EPs, however all of them are statistically significant. There are default parameters provided for both algorithms and some preliminary research shows, that except minimal growth rate for maximal frequent itemsets algorithms the other parameters have limited impact on set of discovered EPs. In Fig. 4 there is a window for choosing algorithm discovering EPs and setting the appropriate parameters and in Fig 5 there is presented set of EPs revealed by algorithm using maximal frequent itemsets.

EPs enable data classification for which CAEP (Classification by Aggregating Emerging Patterns) algorithm [Dong, 1999a] is applied. For this classifier set of EPs discovered by maximal frequent itemsets algorithms gives usually slightly better classification. On the other hand algorithm using decision tree produces more compact and more significant knowledge, which probably can be more interesting for expert trying to update his/her knowledge on the analyzed problem.

Support Vector Machines is yet another methodology being integrated to ARES System. Currently there are attempts to expose results of classification of data done with this approach. Only credibility coefficients calculated for each object from information system are available from ARES System now, but their calculations involve the classification itself. The credibility coefficients are presented in the following section.



no.	Target Class (...	Emerging Pattern	Growth Rate	Target Support	Rest Support
13	Iris-setosa	sepalwidth >= 3.15, sepalwidth < 3.95	3.66667	66 %	18 %
14	Iris-setosa	sepalwidth >= 3.15, sepalwidth < 4.15	3.88889	70 %	18 %
15	Iris-setosa	sepalwidth >= 3.45, sepalwidth < 4.3	13.3333	40 %	3 %
16	Iris-setosa	petallength < 1.45	+∞	46 %	0 %
17	Iris-setosa	petallength < 1.55	+∞	74 %	0 %
18	Iris-setosa	petallength < 1.65	+∞	88 %	0 %
19	Iris-setosa	petallength < 1.8	+∞	96 %	0 %
20	Iris-setosa	petallength >= 1.45, petallength < 1.65	+∞	42 %	0 %
21	Iris-setosa	petallength >= 1.45, petallength < 1.8	+∞	50 %	0 %
22	Iris-setosa	petalwidth < 0.25	+∞	68 %	0 %
23	Iris-setosa	petalwidth < 0.35	+∞	82 %	0 %
24	Iris-setosa	petalwidth < 0.45	+∞	96 %	0 %
25	Iris-versicolor	sepallength >= 5.15, sepallength < 6.15	2.5	60 %	24 %
26	Iris-versicolor	sepallength >= 5.15, sepallength < 6.35	2.1875	70 %	32 %
27	Iris-versicolor	sepallength >= 5.35, sepallength < 5.95	2.625	42 %	16 %
28	Iris-versicolor	sepallength >= 5.45, sepallength < 6.95	2.04762	86 %	42 %
29	Iris-versicolor	sepalwidth < 2.75	3.5	42 %	12 %
30	Iris-versicolor	sepalwidth < 2.85	2.7	54 %	20 %

Figure 5. Emerging patterns produced by algorithm using maximal frequent itemsets

Credibility Coefficients

Calculations of credibility coefficients is a unique feature of ARES System. A credibility coefficient is a heuristic measure, which assesses typicality of a given object in respect to other objects of information system. Value of credibility coefficient ranges from 0 to 1 and lower values denote worse credibility. The concept of credibility coefficients is based on assumption that majority of data is correct. Minority of data can be incorrect or corrupted. The goal of credibility coefficients is to identify this minority by applying different approaches.

Currently in ARES System there is a number of algorithms for calculations of credibility coefficients based on the following concepts:

- Approximation of rough set classes
- Statistics of attribute values
- Hybrid of previous two
- Frequent set
- Extracted Rules
- Voting Classifier (CAEP)
- Support Vector Machines
- Multi Credibility Coefficient

The first five algorithms for calculations of credibility coefficients belong to the original version of ARES System and mostly exploit concepts of rough set theory. They were described elsewhere in details [Podraza 2007] [Podraza, 2006b] . Credibility coefficient based on voting classifier was incorporated as a part of KTDA system. It

takes into account the real classification of the object and vector of weights of votes for classification determined by the classifier. Any voting classifier outcome can be utilized - CAEP and SVM classifiers are used in ARES System.

Values of different credibility coefficients are incomparable, although all belong to interval <0; 1>. There was proposed a new kind of credibility coefficient, namely ordinal credibility coefficient. Ordinal credibility coefficient is associated with any arbitrary chosen "normal" credibility coefficient, whose values are data for the former one. Ordinal credibility coefficient expresses the relative amount of records with credibility coefficients less or equal to the credibility coefficient for this record. Important feature of ordinal credibility coefficient is fact that its and its input counterpart's values introduce the same ordering of data set objects. In other words, ordinal credibility coefficient of a given record defines its relative position within the data set according to ranking given by the value of the basic credibility coefficient. Multi Credibility Coefficient method [Podraza 2007] combines a number of ordinal credibility coefficients to obtain an aggregate outcome. The resulting value is (weighted) average of aggregated ordinal credibility coefficients.

Credibility coefficients are supposed to reveal "improper" data. Quite often it may be extremely important to focus user attention on such exceptional cases. For instance in medical application the case supporting known rules and procedures can be proceed routinely, while exceptions may require extra check-up and treatment. And in practical data analysis exceptions may appear more interesting then the rules themselves.

Conclusion

The paper presents ARES System functional capabilities. All typical stages of data exploration based on the rough set theory can be performed and presented with support of ARES System. The system has been extended by Emerging Patterns approach and Support Vector Machines methodology.

ARES System has its unique characteristics for discovering non-typical objects in information system. Credibility coefficients are used to evaluate object's measure of typicality in respect to the rest of the information systems. Many algorithms for evaluating credibility coefficients are offered.

ARES System was designed to be used in medical application. Medicine and natural sciences appear to be often interested in exceptions more than in rules – a patient, who reacts exceptionally to a routine treatment causes the highest concern of physicians. ARES System's unique feature to recognize exceptional cases by employing credibility coefficients seems valuable to medicine and other natural sciences but, generally, it can be used on any kind of data, e.g. for engineering purposes.

A multi-document architecture of the ARES System allows for detailed analysis of the data exploration process, what makes the system a perfect and easy to use learning tool.

The ARES System has been implemented in Java and is portable. Its architecture enables permanent development by adding new items with appropriate algorithms to the documents presented and processed by the system. The module structure of ARES System makes its development quite obvious – new functional items inherit structure features of the system.

Although ARES System has been designed to allow its permanent development. It is planned to implement ARES System as Service Oriented Architecture (SOA). The server part will contain all services, very often mutually independent, responsible for steps of data analysis. A number of client programs, tailored to user needs, will call the services. This approach should support developing the system by adding new methodologies, updating existing algorithms and testing parts of the system without interfering with its working part. For applications it should be very attractive to compare results of data analysis performed by different approaches.

Bibliography

[Podraza 2005] Podraza R., Walkiewicz M., A. Dominik A.: Credibility Coefficients in ARES Rough Set Exploration System. Proc. 10th Int. Conf. on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, RSFDGrC 2005, Regina, Canada, Lecture Notes in Artificial Intelligence, LNAI 3642, Part II. Springer-Verlag, Berlin Heidelberg New York (2005) 29-38.

[Pawlak, 1991] Pawlak Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer, (1991).

[Podraza, 2007] Podraza R., Tomaszewski K.: Ordinal credibility coefficient - a new approach in the data credibility analysis. In A. An, J. Stefanowski, S. Ramanna, C. J. Butz, W. Pedrycz, G. Wang (Eds.), Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Proc. 11th Int. Conf., RSFDGrC 2007, Toronto, Canada, Lecture Notes in Artificial Intelligence LNAI 4482, Springer-Verlag Berlin Heidelberg (2007) 190-198.

[Podraza, 2006] Podraza R., Dominik A.: Problem of Data Reliability in Decision Tables. Int. J. of Information Technology and Intelligent Computing (IT&IC), Vol. 1 No. 1, (2006) 103-112.

[Grzymala, 2005] Grzymała-Busse J.W.: Rule Induction. In Data Mining and Knowledge Discovery Handbook. Maimon O., Rokach L. (Eds.), Springer Science + Business Media Inc. (2005) 277-295.

[Dong, 1999] Dong G., Li J.: Efficient Mining of Emerging Patterns: Discovering Trends and Differences. Proc. of the SIGKDD (5th ACM Int. Conf. on Knowledge Discovery and Data Mining), San Diego, USA, (1999) 43-52.

[Schmilovici, 2005] Schmilovici A.: Support Vector Machines. In Data Mining and Knowledge Discovery Handbook. Maimon O., Rokach L. (Eds.), Springer Science + Business Media Inc. (2005) 257-274.

[Podraza, 2006a] Podraza R., Tomaszewski K.: KTDA: Emerging Patterns Based Data Analysis System. Annales UMCS, Informatica, AI, Vol.4, Lublin, Poland, (2006) 279-290.

[Boulesteix, 2003] Boulesteix A.-L., Tutz G.: A Framework to Discover Emerging Patterns for Application in Microarray Data. Institute for Statistics, Ludwig-Maximilian-Universitat, Munich (2003).

[Dong, 1999a] Dong G., Zhang X., Wong L., Li J.: CAEP: Classification by Aggregating Emerging Patterns. Proc. of 2nd Int. Conf. on Discovery Science, Tokyo, Japan, (1999) 30-42.

[Podraza, 2006b] Podraza R., Dominik A.: Credibility coefficients for objects of rough sets. Studia Informatica, No. 1/2 (7) (2006) 93-104.

Authors' Information

Roman Podraza – Institute of Computer Science, Warsaw University of Technology; Nowowiejska 15/19, 00-665 Warsaw, Poland ; e-mail: R.Podraza@ii.pw.edu.pl

Mariusz Kalinowski – Institute of Computer Science, Warsaw University of Technology; Nowowiejska 15/19, 00-665 Warsaw, Poland ; e-mail: M.M.Kalinowski@elka.pw.edu.pl