Third International Conference
Modelling and Development of Intelligent Systems
October 10 - 12, 2013
“Lucian Blaga” University
Sibiu - Romania

Methodological Framework for Creating a Workflow Model when Processing Data Research

Alexandra-Mihaela Pop, Ioan Pop

Abstract

In this article we present a methodological approach for creating workflow models. Using a workflow model for data processing research can be considered a tool for quantitative management into project. The advantage of the approach of a workflow model in the research is that the data collected by investigating a large population studied, is processed automatically, dynamically and executed with real-time machine learning methods. In practice, this approach leads to the enrichment of the “Project Management Body of Knowledge”. Also the methodological framework can serve as a tool to initiate and train students interested in developing research projects. The article describes a scenario for creating a workflow model for communication management.

1 Introduction

Processing data from different research processes involves a computationally intensive, statistical analysis and interpretation techniques based on machine learning methods execution. To speed up and automate data processing it is necessary to create workflows that support the researcher or expert. Workflows can be created with different toolboxes specific data processing tasks. There are toolboxes for creating workflows for business - business workflows - but also for scientific research - scientific workflows. Scientific workflows can be composed of steps that follow the stages of a research process such as: acquisition, integration, reduction, analysis, visualization, and publication (e.g. in a shared database) of scientific data [02].

For researchers a scientific workflow is a model that can assure the automatization of data processing by executing the model in a repetitive way. Also, a scientific workflow is an additional support for monitoring the execution of algorithmic methods in real-time.

In the first part of this article we present ways to model the processes researched, a mathematical model of a workflow, and a framework for creating a workflow.

The second part is devoted to using the techniques for creating workflows with the help of the Weka toolbox.

In the last part of the article we present a scenario in which we created a working model of such a workflow. Workflow model execution from the case study helps a researcher focus on the analysis of communication in project management. Based on the results of workflow execution, the researcher can draw interpretations motivate members by communicating in the project.

2 Modeling as a tool for manipulation the applicative and constructive entities

2.1 Ways of modeling

Natural or artificial processes are the object study of people concerned with increasingly discovering reality. In order to be understood and interpreted correctly these processes are represented by different models. As abstractions of essential attributes, models are characterized by a certain fidelity of the processes but the most important thing is that they help by simulating their execution in obtaining reliable solutions obtained after processing.

Modeling is done in various forms in relation to the methods used in the field where it is produced. For example, in architecture iconic designs is used, while in economics symbolic modeling is used (through mathematical models). All modeling techniques are based on the mechanism of abstraction of things, phenomena and processes. Thus, obtaining an execution flow model for data processing is a good method for handling, automation and optimization of these processes. Methodologically speaking, designing a workflow model involves identifying the basic elements of the new workflow, assess and document key characteristics of process modeling, i.e. the following steps [04]:

- Outline the workflow;
- Detailing the various levels of work;
- Evaluation and verification of sustainability at every level;
- Review, or move to the next level of detail;
- Iteration elements;
- Review on a larger sample.

2.2 Mathematical model of the workflow

As a model of computation, a workflow can be abstracted by a mathematical representation as a graph as a set of pairs actor-connection, where the actor can be: a job, task or step work, and the connection is an arc routed [02]. Computation model on the workflow has the following notation: at $\boldsymbol{w}$ graph is associated with a set of input parameters $\boldsymbol{p}$, a set of input data $\boldsymbol{x}$ and a set of output data $\boldsymbol{y}$. Therefore, we represent the model as a workflow application form

$\boldsymbol{M}\colon \boldsymbol{W} \times \boldsymbol{P} \times \boldsymbol{X} \rightarrow \boldsymbol{Y}$, where $\boldsymbol{w} \in \boldsymbol{W}, \boldsymbol{p} \in \boldsymbol{P}, \boldsymbol{x} \in \boldsymbol{X}$ and $\boldsymbol{y} \in \boldsymbol{Y}$.

$\boldsymbol{M}$ application is defined as $\boldsymbol{y} = \boldsymbol{M}\,(\boldsymbol{Wp}\,(\boldsymbol{x}))$, i.e. any $\boldsymbol{w}$ workflow of $\boldsymbol{W}$ for proper parameter set to $\boldsymbol{p}$ and input $\boldsymbol{x}$, workflow determines an output $\boldsymbol{y}$. This graph model is found in the Kepler system, a system for creating scientific workflows [11].

2.3 A framework for workflow model

A framework based on a workflow model can support the process management task be it social, economic or scientific. The management of the scientific processes can be assisted by framework which is modelled through a scientific workflow.

Scientific workflows have specific operations research and are planned during the design flow. Execution is conducted at runtime; the user specifies data processing operations flow while dependencies are specified by the workflow designer. Typically, workflows are represented visually by the use of block diagrams, or by specifying them by means of a specific programming language [06].

The researcher utilizes a workflow as a deliverable in his project as a recipe to automate, document, and make possible a scientific process repeatability. Thus the scientific workflow has a life cycle like any other artefact from the project.

The life cycle of scientific workflow route is feasible and has associated phases as: development, implementation and execution of scientific workflows. These phases are largely an endorsement by the workflow systems, systems that have methods and techniques of data mining.

Table 1 shows the key elements underlying the design of a workflow model to a process of analysis based on a relationship of communication.

Table 1. Phases of a design methodology for workflow model.
(adapted from A. Sharp and P. McDermott, *Workflow Modeling)*

1) Establish process context, scope and goals	**2) Understand as-is process-workflow and other enables**	**3)Define to-be process characteristics and requirements**
•***Identify related processes*** - identify and link activities - 1:1 links are in same process - draw Overall Process Map •***Clarify target process' scope*** - triggering event, ~5+/- processes, result for each stakeholder, cases/variations •***Clarify as-is process elements*** - functional areas - actors and responsibilities - systems and mechanisms •***Assess as-is process by stakeholder*** (initial) - also specify context and consequences of inaction •***Specify to-process goals*** - subjective and objective •***Specify performance metrics*** - customer-focused outcomes, not internal task efficiency	•***Organize and initiate session*** - staff and management plus external stakeholders - review scope, issues, goals - review ground rules •***Build as-is swimlane diagram*** - one case and path at a time - 1)"Who gets it next?" 2)"How does it get there?" 3)"Who *really* gets it next?" •***Check each step - 5 questions*** 1) again "How does it get there?" 2)"No mushy verbs?" 3)"All triggers shown?" 4)"All participant actors shown?" 5)"All outputs shown?" •***Model other process cases*** - create new diagram, or use original case as a starting point •***Add additional levels of detail*** - only if necessary	•***Assess as-is process by enabler*** (final assessment) - using as-is diagram as a guide - helps us take a holistic view •***Decide on approach*** (abandon, outsource, leave as-is, improve, or redesigne) •***Conduct challenge session*** - challenges hidden assumptions, generates creative ideas - helps us "*think out of the box*" •***Eliminate infeasible ideas*** (cost, legal, resources, impact, ...) •***Assess improvement ideas by enabler*** - helps us *avoid unanticipated consequences* - builds requirements document •***Lay out to-be workflow*** - handoff level first, then milestone and task levels

Three categories of tasks are shown in Table 1, which underline building a workflow model: 1) Entry in the context of the research process by clarifying the scope, objectives and performance metrics, 2) Understanding the workflow as the research process the task of organizing, building work steps, creating charts and other details, 3) Defining the process and requirements for workflow modeling.

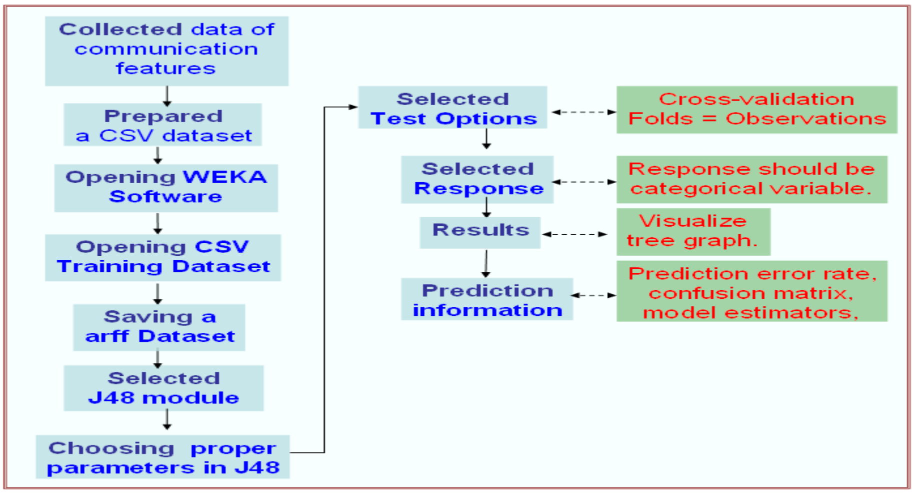

Collected data of communication features
Prepared a CSV dataset
Opening WEKA Software
Opening CSV Training Dataset
Saving a arff Dataset
Selected J48 module
Choosing proper parameters in J48
Selected Test Options
Cross-validation Folds = Observations
Selected Response
Response should be categorical variable.
Results
Visualize tree graph.
Prediction information
Prediction error rate, confusion matrix, model estimators.

Figure 1: Workflow model for creating and processing a communication relationship.

A scientific workflow can be created with the Weka toolbox. It has even been called "Knowledge Flow" environment. This application has the possibility to create a flow of data processing to get a model to be represented as accurately as possible and help with the simulation of a scientific process. Moreover, it contributes to the implementation of the so-called quantitative management.

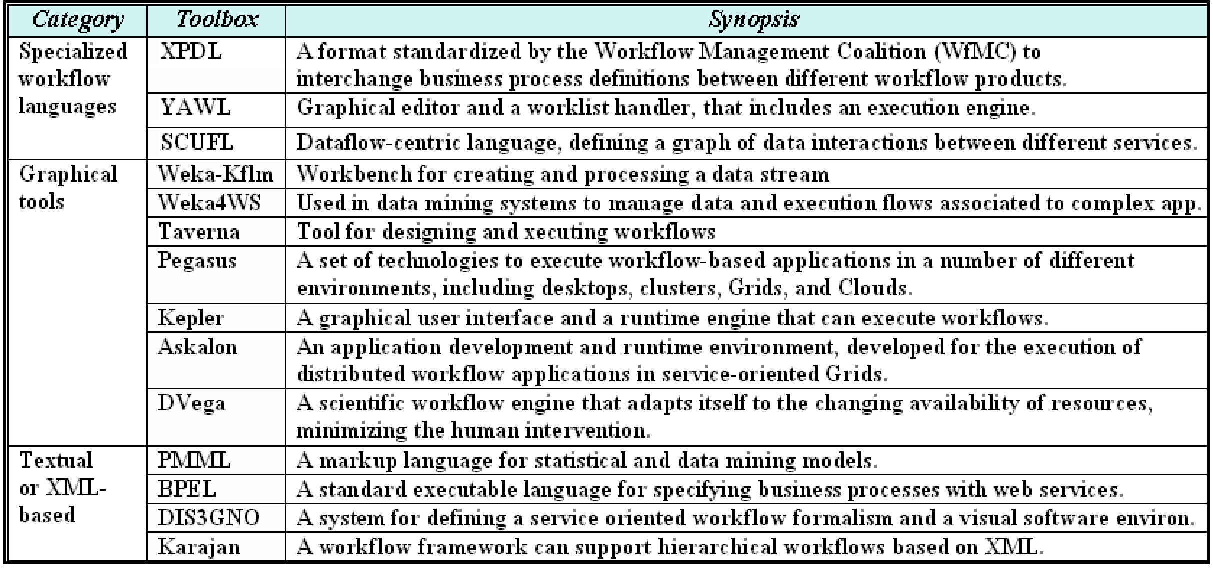
A flow pattern is illustrated in Figure 1. This model simulates the processing of instances of a relationship with attributes of a process of communication.

In the model in Figure1 we have followed the design of a workflow designed for processing a communication relationship. That is, we structured and framed the research process, we set semantics of tasks and next we designed the flow model (model that reproduces the process).

3 Modeling technique by Weka

There are several toolboxes that can be used for the creation of workflows as can be seen in summary of Table 2.

Tabel 2: Categories of toolboxes for the management and execution of workflows [03][08].



Category	*Toolbox*	*Synopsis*
Specialized workflow languages	XPDL	A format standardized by the Workflow Management Coalition (WfMC) to interchange business process definitions between different workflow products.
	YAWL	Graphical editor and a worklist handler, that includes an execution engine.
	SCUFL	Dataflow-centric language, defining a graph of data interactions between different services.
Graphical tools	Weka-Kflm	Workbench for creating and processing a data stream
	Weka4WS	Used in data mining systems to manage data and execution flows associated to complex app.
	Taverna	Tool for designing and xecuting workflows
	Pegasus	A set of technologies to execute workflow-based applications in a number of different environments, including desktops, clusters, Grids, and Clouds.
	Kepler	A graphical user interface and a runtime engine that can execute workflows.
	Askalon	An application development and runtime environment, developed for the execution of distributed workflow applications in service-oriented Grids.
	DVega	A scientific workflow engine that adapts itself to the changing availability of resources, minimizing the human intervention.
Textual or XML-based	PMML	A markup language for statistical and data mining models.
	BPEL	A standard executable language for specifying business processes with web services.
	DIS3GNO	A system for defining a service oriented workflow formalism and a visual software environ.
	Karajan	A workflow framework can support hierarchical workflows based on XML.

A good toolbox to achieve workflows for different areas such as the scientific, medical, social, economic is Weka. Weka Workbench has an interface for creating and processing a data stream called KnowledgeFlow. With KnowledgeFlow tool we can create frameworks planned to automate and execute a model or even a Scientific Workflow Life Cycle. The KnowledgeFlow supports Life Cycle phases including design and workflow composition, workflow resource planning, execution workflow execution analysis, visualization and dissemination.

A workflow created in Weka KnowledgeFlow has the following data processing and relationship attributes as input: a layout style of intuitive data flow, processing data in batches or incrementally, processing multiple batches or streams in parallel (each separate stream runs in its own thread) chains of data filters, prospects for models produced by classifiers for each cross-validation, visualization performance incremental classifiers during processing (classification accuracy, RMS error, predictions, etc.) facility to allow easy addition of new components to KnowledgeFlow (plug-ins).

When creating a workflow model, specifically simulate a process that complies with the majority, the next steps (with sub-steps) to build a graph with nodes and connectors: 1) the addition of nodes needed (add new node; add corresponding data source node; assignment class, and adding CrossValidationFoldmaker node; add the node classifier, adding performance evaluation node, add visualization, etc..); 2) connecting nodes (connecting the two by two and chaining nodes) ; 3) execution of the startup process of the DataSource node "Start loading", 4) viewing the results (if the execution was done correctly, the results will be presented in a separate window).

4 Case Study: Project Communication Management

We present a detailed case study to illustrate the concepts discussed in the previous sections. More concretely we chose a workflow for processing communication data. The data is structured in a relationship with instances of communication.

Each instance contains values motivating communication characteristics. The communication relationship has seven attributes: five attributes are characteristics of communication motivating (satisfaction, trust, openness, listening, encouraging), and two are auxiliary attributes (role communicator class project and communication). Auxiliary attribute called ClasaCom is one with binary values and aims to assign each instance of communication value ("yes" above average characteristics of motivation and "no" otherwise). ClasaCom attribute is important to train classification or clustering methods in the model created with Weka.

We have created a workflow with KnowledgeFlow tool for researcher studying communication projects.

This workflow involves creating a model of analysis and interpretation of communication motivation for the a portfolio of IT projects. Here there are a series of specific problems which are not found in other types of research workflows. In our view, the workflow is a directed graph, ie a flow of execution of several data processing tasks to build a model of motivating communication. On the other hand in the model we execute the workflow in order to deliver results for the interpretation of communication motivation.

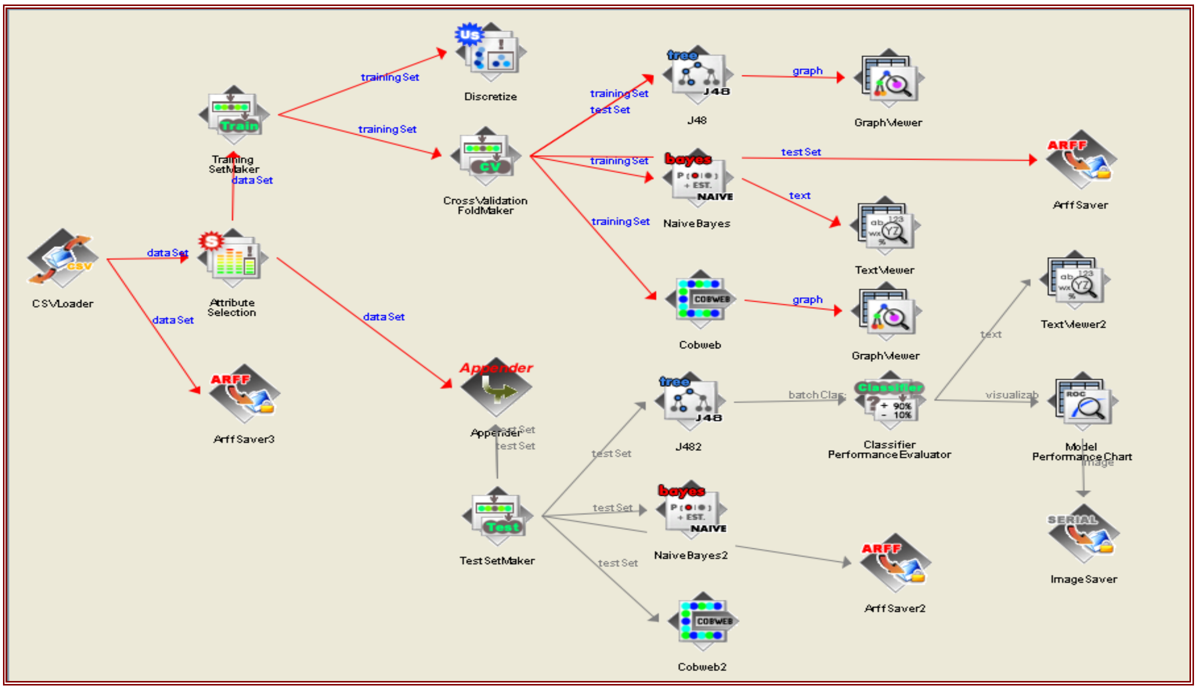


Figure 3: Workflow for data processing of motivating communication.

Resource input workflow is a communication relationship populated with values recorded by interviewing different stakeholders in software development projects. Attributes of a communication instance have characteristic values of motivation, measured on a Likert scale from 1 to 5, values that describe motivating communication management projects at different levels of communication. Task by the stream processing model does not contain complex computations, but the application of statistical techniques lead to new interpretation motivational attributes of communication quality in IT projects. Also, the flow of execution classifications / clustering and filtering leads to results that can be viewed through textual tables, graphs and diagrams necessary for the qualitative interpretations of motivation communication projects.

During flow model different backups, charts, graphs and data archiving management are made. Such a workflow that focuses on the detailed running Weka system, communication researcher is left to deal with the results of the execution model, the researcher end user workflow.

Such a worflow is illustrated in Figure 3.

Mapping Workflows to Resources – shall be made the start node of the workflow, where the input is a spreadsheet of the type CSV, which node 2 is transformed into a dataset, file type ARFF. Once the data resource has been assigned to the Workflow it is available for any type of connection allowed in Workflow processing. Resource input is a relation that contains instances of communication and motivation attributes values of communication, as described above.

Workflow Execution – can be done in two ways: automatically or bach. Workflow in Figure 3 runs bach researcher monitored. During processing flow runs: attribute selection, filtering data test drive data execution algorithms, running algorithms on test data, etc.

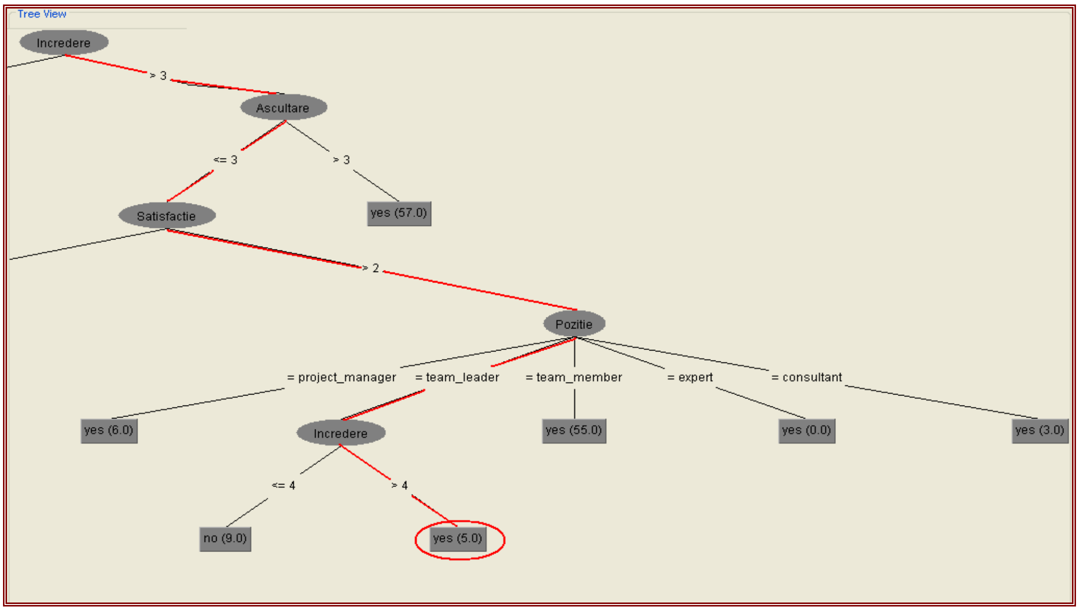
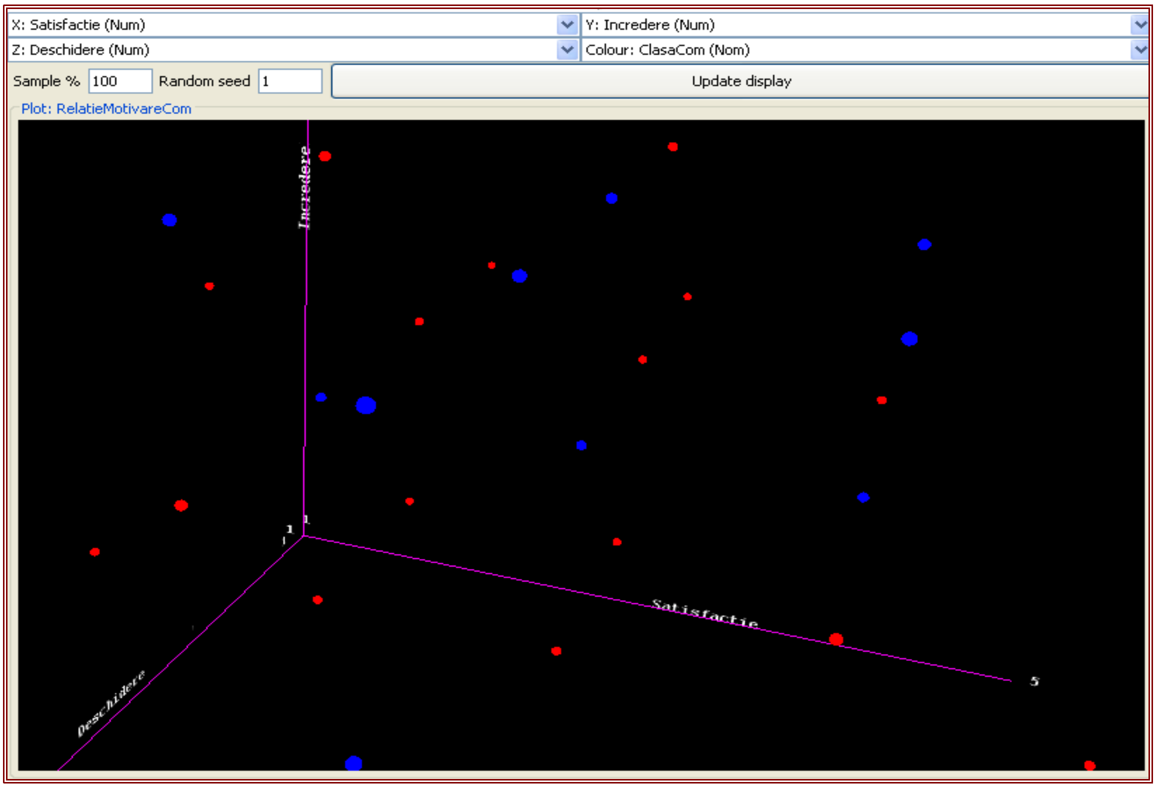


Figure 4: Part of J48 graph wiew with the branchs of the decision tree.

A result from procesing the data, after a clssifying algorythm from the clas DC45 has been applied, is a J48 graph presented in figure 4 and a 3D chart of the motivation characteristics of communication presented in Figure 5. In the graph from Figure 4 a trail is shown which suggests a decision rule for obtaining optimal motivation from communication in projects. In the chart from Figure 5, which is represented in 3D, we have the distribution and the characteristics of motivation in the communication process by satisfaction, trust, an opening.



X: Satisfactie (Num)

Y: Incredere (Num)

Z: Deschidere (Num)

Colour: ClasaCom (Nom)

Sample % 100

Random seed 1

Plot: RelatieMotivareCom

Figure 5: A 3D reprsentation of the communication characteristics of motivation.

Workflow Reuse – this workflow is saved as a .kfml file and an be used later on other data that is being collected. Also results of the workflow execution can be saved which in turn can be used for interpretations, comparisons and analysis.

5 Conclusions

Since the area of research is widening and research time spent is becoming more precious the working mechanism proposed in this article is a good design for more efficient and faster processing research data.

Methodological considerations suggested us to create efficient workflows is a framework for building workflows. Workflows automatize iterative tasks of data processing, so that researchers can better focus on the management experiment, and in this way they can more effectively manage the research process.

Workflows created with Weka are easy ways to process data using statistical techniques and methods of machine learnig. In developing a workflow there are various posiblities of creating important views of the results obtained after the data processing takes places.

Doing research work supported by workflows, encourage reuse of future results prelurare by automating a process, both within a project and a portfolio of projects.

The workflow environment presented can be extrapolated and used in distributed process research data, in grid computing and in web service orietd. Moreover, Weka has a working version Weka4WS which can design the workflow-oriented Web services (Web Service-Oriented).

References

[1] I. H. Witten, E. Frank, M. A. Hall, *DataMining: Practical Machine Learning Tools and Techniques*, Elsevier, Burlington, USA, 2011.

[2] B. Ludascher et al, Scientific Process Automation and Workflow Management, in *Scientific Data Management: Challenges, Existing Technology, and Deployment*, Computational Science Series, chapter 13. Chapman & Hall/CRC, 2009.

[3] D. Talia, Workflow Systems for Science: Concepts and Tools, *ISRN Software Engineering*, vol. 2013, Article ID 404525, 15 pages, 2013. doi:10.1155/2013/404525.

[4] A. Sharp, P. McDermott, *Workflow Modeling: Tools for Process Improvement and Applications Development*, Second Edition, Artech House, Norwood, MA 02062, 2009.

[5] B. Ludascher, et al., Scientific Workflow Management and the Kepler System, in *Concurrency and Computation: Practice & Experience*, 18(10):1039–1065, 2006.

[6] R. Prodan, T. Fahringer, *Grid Computing: Experiment Management, Tool Integration, and ScientificWorkflows*, Springer-Verlag, Berlin Heidelberg, 2007.

[7] P. Sharma, A. Rajavat, Analysis and Design of Service-Oriented Framework for Executing Data Mining Services on Grids, *International Journal of Advanced Research in Computer Science and Software Engineering* 3(3), March - 2013, pp.398-402.

[8] http://en.wikipedia.org/wiki/, *workflow system webpages* accessed at jan. 2013.

[9] WekaDoc, http://weka.sourceforge.net/*wekadoc*/ , accessed at mar. 2013.

[10] http://weka.wikispaces.com/*Exporting+Charts+from+the+Knowledge+Flow*, accessed at jun. 2013.

[11] http://www.cs.waikato.ac.nz/~fracpete/downloads/#kepler, accessed at aug. 2013.

[12] W. Tan, M. Zhou, *Business and Scientific Workflows: A Web Service-Oriented Approach*, published by JohnWiley & Sons, Inc., Hoboken, New Jersey, 2013.

[13]F. Fernandez et al., Assisting Data Mining through Automated Planning, *in MLDM 2009, LNAI 5632*, pp. 760–774, Springer-Verlag Berlin Heidelberg, 2009.

Alexandra-Mihaela Pop
University "Lucian Blaga" of Sibiu
Faculty of Engineering
Department of Industrial Engineering and Management
Str. Emil Cioran, Nr.4, Sibiu, 550025
ROMANIA
E-mail: alexandrapop_6@yahoo.com

Ioan POP
University "Lucian Blaga" of Sibiu
Faculty of Sciences
Department of Mathematics and Informatics
Str.Dr.I.Ratiu, Nr.5-7, Sibiu, 550012, ROMANIA
E-mail: ioan.pop@ulbsibiu.ro