Data mining based collaborative analysis of microarray data

Georgia Tsiliki and Sophia Kossida
Bioinformatics and Medical Informatics Team
Biomedical Research Foundation
Academy of Athens, Greece
Email: gtsiliki@bioacademy.gr, skossida@bioacademy.gr

Natalja Friesen and Stefan Rueping
Knowledge Discovery Group
Fraunhofer Institute IAIS
Sankt Augustin, Germany
Email: natalja.friesen@iais.fraunhofer.de, stefan.rueping@iais.fraunhofer.de

Manolis Tzagarakis and Nikos Karacapilidis
Computer Technology Institute and Press "Diophantus",
Rio Patras, Greece
Email: tzagara@upatras.gr, nikos@mech.upatras.gr

Abstract—Biomedical research has recently seen a vast growth in publicly and instantly available information, which are often complementary or overlapping. As the available resources become more specialized, there is a growing need for multidisciplinary collaborations between biomedical researchers to address complex research questions. We present an application of a data-mining algorithm to gene-expression data in a collaborative decision-making support environment, as a typical example of how multidisciplinary researchers can collaborate in analyzing and biologically interpreting gene-expression microarray data. Through the proposed approach, researchers can easily decide about which data repositories should be considered, analyze the algorithmic results, discuss the weaknesses of the patterns identified, and set up new iterations of the data mining algorithm by defining other descriptive attributes or integrating other relevant data.

Keywords - decision-making; collaboration; gene-expression data; text-mining

I. INTRODUCTION

Recently, biomedical research has become increasingly interdisciplinary and collaborative in nature [7], [13], [19]. A vast growth in publicly available biomedical resources, including multiple types of data sets and databases, analysis tools, and statistical methodologies are now available. Furthermore, a major advance is that now researchers have access to complementary views of a single organism by analyzing multiple types of data, including whole genome sequencing, expression profiling and other high-throughput experiments [25]. At the same time, applications which link together bioinformatic tools and databases have emerged, showing the way to easily visualize and analyze biomedical data. For instance, BioGRID [24], BND [6], BIOZON [6], BioMart [12] and its cancer specialized version IntOGen, are repositories which store readily combined data sets and provide platforms to easily visualize data. Similarly, the GenePattern platform provides access to more than 180 tools for genomic analysis to enable reproducible in silico research.

However, extended functionality and integration with other, more advanced, systems is often required, especially as the number of related Web services is constantly increasing. Some initial efforts have been conducted to facilitate the collaboration, data sharing and decision-making between scientists by providing a common platform environment where users can share their common interests in a forum-like manner. A well known example of such a virtual community is myExperiment, an online research environment that supports the social sharing of bioinformatics workflows, i.e., procedures consisting of a series of computational tasks, which can then be reused to their specific requirements [9]. The Galaxy Project offers a web-based platform allowing researchers to perform and share analyses. BioCatalogue [5], a Web based service registry, allows users to annotate and comment on the available services in order to assist them in identifying suitable services.

Supporting collaboration in biomedical research is still considered a challenging task. Researchers need to efficiently and effectively collaborate and make decisions by appropriately assembling and analyzing enormous volumes of complex multi-faceted data residing in different sources, without having to deal with the problem of information overload. In such processes the integration of other activities such as data mining, i.e. the process of pattern and dependencies discovery within large data sets, are beneficiary exhibiting the need of a holistic approach towards the integration of collaboration, decision making and data mining services.

The approach described in this paper enables researchers to fully embed data mining in a collaborative data analysis and decision making process. Our goal is to tightly integrate data mining and collaboration support services in an interactive environment. This approach is being implemented in the context of the EU funded Dicode research project (http://dicode-project.eu/), which aims to facilitate and augment collaboration and decision making in diverse data-intensive and cognitively complex settings. The remainder of this paper briefly introduces the web services offered in Dicode, describes in detail a particular interactive data mining service implemented in the context of this project, and - through a concrete example - illustrates its seamless integration with collaboration support services.

II. THE DICODE APPROACH

Dicode provides a novel Web-based environment with advanced data mining and collaboration support services. Cen-
tral to the Dicode approach is the concept of workbench. Workbenches enable the seamless integration of various data analysis and collaborative decision making support services (Fig. 1). Interoperability of services is achieved from both a technical and a conceptual point of view. In this regard, semantics techniques have been exploited to define an ontological framework for capturing and representing the diverse stakeholder and services perspectives. A Dicode workbench can be configured by end users as far as the services to be included are concerned (according to the needs of the particular context and problem under consideration). A widget-based approach has been adopted to offer the relevant services. The central widget of Fig. 1 provides access to collaboration and decision making support services. On the right, the search service widget allows users to locate new services. Widgets on the left are customizable by the user and concern data mining services. Services provided through the Dicode workbench are:

- **Data acquisition services**, which capture the decisive tractable information existing in diverse data sources and formats. Particular attention is given to web resources and the development of the associated spider components (web crawler).
- **Data pre-processing services**, which efficiently manipulate raw data before their storage to the foreseen solution. Transformation of different kinds of documents into a canonical form, structuring of documents, data cleansing, as well as language detection and linguistic annotations, are some of the functionalities foreseen in this category of services.
- **Data mining services**, which are built on top of cloud infrastructure, high performance full text search, data indexing, classification and clustering, directed data filtering, and meaningful data aggregation. Text mining techniques such as named entity recognition, relation extraction and opinion mining are employed to aid the extraction of valuable semantic information from unstructured texts. Elaborated intelligent data mining techniques include local pattern mining, similarity learning, and graph mining.
- **Collaboration support services**, which assist users’ collaboration through adaptive workspaces, handle the visualization of the data mining services outcomes (through alternative and dedicated data visualization schemas), and provide a workflow engine enabling a series of actions for the appropriate handling per case data.
- **Decision making support services**, which enhance both individual and group sense- and decision-making by supporting stakeholders in identifying, retrieving and arguing about relevant information or knowledge, whilst providing them with appropriate notifications and recommendations given their preferences, competences, expertise etc., exploit the reasoning capabilities of humans.

Focusing on a particular data mining service developed in the context of the Dicode project, namely Subgroup Discovery (SD), and exploiting a particular collaboration case concerning analysis and interpretation of gene expression microarray data, the next sections discuss in detail the proposed interoperation of data mining and collaboration support services.

### III. INTERACTIVE SUBGROUP DISCOVERY SERVICE FOR FUNCTIONAL INTERPRETATION OF GENE EXPRESSION DATA

One of the most common tasks in gene expression analysis is the estimation of any available structure or patterns in the data [10], [26]. For example, a common task is to determine whether or not gene expression data can reliably identify or classify different types of a disease. Researchers normally apply classification, clustering and partition algorithms [23] to estimate the existing structure, if any. Most of those algorithms are already implemented in popular software and are well presented. Other algorithms which specialize on estimating patterns in the data, though largely applied on literature search and categorization of database annotations, are the text-mining algorithms [1], [8]. The ultimate goal of text mining concerns the discovery of new, previously unknown information, by automatic text resources processing.

In terms of the interpretation of results derived from the case discussed in this article, the main purpose of a typical microarray experiment is to find a molecular explanation for a given macroscopic observation. The most common methods are based on the ‘functional enrichment’ that first selects genes of interest (e.g. genes that are significantly over- or under-expressed when two classes of experiments are compared) and then includes external sources of information, such as gene ontologies and pathways databases, to translate the set of genes into interpretable biological knowledge.

Regarding both the analysis and interpretation of results, there is an established need for easy-to-use services that support any researcher to make use of popular existing methodologies which can be also applied in an iterative fashion according to the received feedback. Moreover, it is important to enable the user to share his results with other researches that are working on the similar problems and to discuss the weaknesses of the identified patterns.

We address these problems by providing the user with data mining services that are included in a collaborative workbench. Fig. 2 shows a scenario of using the SD service in Dicode workbench as an interactive process. First, the differentially expressed genes that were identified by the user are automatically enriched using Gene Ontology (GO) information. Next, the algorithm finds the most interesting subgroups that functionally interpret differentially expressed genes. The discovered subgroups are presented in a collaborative workbench as one piece of knowledge and can be interpreted and discussed by experts in the overall discussion process. According to the feedback the user can set up a new iteration of the algorithm by specifying undesired attributes or non-interesting subgroups. The suggested procedure could handle multiple-platform gene-expression data, since it can project the biological significance of the data set relative to the disease at hand and irrespectively of the data scale. The
importance of the approach proposed here will be made clearer in the next two sections, where Dicode collaboration support services are discussed in more detail.

The user interface to the SD service is presented in Fig. 3. The parameters serve in both controlling the complexity of the output, such as the number of retrieved subgroups, as well as in defining a particular knowledge database that is used for interpretation of genes. Currently only the three components of the GO are integrated in the service, nevertheless we aim to include further knowledge databases in the future, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [14].

A. Subgroup Discovery

SD is the task of finding patterns that describe subsets of a data set highly correlated relative to a target attribute. This is a popular approach for identifying interesting patterns in data, because it combines statistical significance with an understandable representation of patterns. For example, in a group of patients that did or did not respond to specific treatment, an interesting subgroup may be that patients who are older than 60 years and do not suffer from high blood pressure respond much better to the treatment than the average.

Let \( A_1, \ldots, A_d \) be a set of attributes, each with a finite domain \( \text{dom}(A_i) = \{ v_{i,1}, \ldots, v_{i,k} \} \) and define \( X = \text{dom}(A_1) \times \ldots \times \text{dom}(A_d) \). Also, let \( Y = \{ 0, 1 \} \). For a fixed data set \( D = \{(x_i, y_i)\}_{i=1}^{n} \subseteq X \times Y \) we define \( p_0 = \frac{1}{n} \sum_i y_i \) and for any subset \( D' \subseteq D \) let \( g(D') = |D'|/n \) and \( p(D') = \frac{1}{|D|} \sum_{(x, y) \in D'} y_i \). \( g(D') \) is called the generality of \( D' \) and \( p(D') \) the expectation of \( Y \) given \( D' \). Also, for a propositional formula \( S \) over \( A_1, \ldots, A_d \) let \( \text{ext}(S) \) be the extension of \( S \) in \( D \), i.e. the set of all \( (x, y) \in D \) for which \( S(x) \) is true. We define \( g(S) = g(\text{ext}(S)) \) and \( p(S) = p(\text{ext}(S)) \). \( S \) is called a subgroup descriptor and \( \text{ext}(S) \) the corresponding subgroup (however, without loss of generality, in the following we will also call \( S \) a subgroup if no confusion is possible).
a) Definition 1:: Given a database $D$ and some $a > 0$, the quality $q$ of a subgroup $S$ is defined as $q(S) = g(S)^a (p(S) - p_0)$

Note that for $a = 0.5$, which is usually used here, this quality is order-equivalent to a Binomial test. The task of SD can now be formally defined as follows:

b) Definition 2:: Given a database $D$ and an integer $k$, the task of SD is to find among all possible subgroups the $k$ subgroups with largest quality values $q$.

Algorithmically, SD ranks possible subgroups according to a quality function $q$, which depends on both the size of the subgroup (in the example above, the number of patients in the study over 60 years without high blood pressure) and the probability of the target attribute in the subgroup (the percentage of patients in the subgroup that responded well to the treatment). That is, subgroup discovery prefers subgroups where the distribution of the target attribute is unusually high, denoting interesting results, but at the same time the subgroups are large denoting that the results are reliable. A popular choice for the quality function is to use the statistical significance of the subgroup as calculated by the binomial test. Given a set of descriptive attributes (e.g. age and diagnosis), the space of possible subgroups consists of all rules that can be formed by conjunctions of attribute-values comparisons. SD tests all possible subgroups and reports the top $k$ subgroups with respect to the given quality function. Efficient algorithms for fast SD exist [11], [15].

However, the power of a subgroup to the user is not the same as the statistical significance of the subgroup. In the context of decision support, for example, the user will prefer to see subgroups containing attributes that are influenceable by the user. This property is called actionability in [18].

In [3] an overview of several techniques for generating interesting subgroups using background knowledge is given, e.g. defining filters on rules and on the subgroup search space. [4] presents approaches and techniques for the refinement and assessment of subgroups, e.g. by visualization of the results. [21] presents the idea of automatically optimizing subgroup sets by learning a ranking function based on user preferences, re-ranking the result set, and feeding back input in the SD process to close the loop. The idea to use gene ontology for functional interpretation of microarray experiments has become popular recently. Many tools exist for this type of analysis, particularly a detailed comparison of fourteen tools is presented in [16]. To the best of our knowledge there are limited approaches that combine data mining and functional interpretation of gene data in an interactive and collaborative way. Such approaches can be proved to be very helpful to researchers without specific knowledge in this challenging task.

IV. EXPERIMENTS

The main contribution of our work is the integration of the well known SD algorithm and the approach for functional enrichment of gene data sets using knowledge databases into a collaborative workbench. Moreover, we provide the user with an easy-to-use interactive service that enables him to improve data analysis results by taking into account a collaborative feedback.

A. Data

In order to validate our service in a realistic scenario we performed iterative SD including feedback from the expert. The data set used in our evaluation was published in [2] where authors study integrating global gene expression and radiation survival parameters to develop a database of responses of cancer cells to ionizing radiation. Particularly the data considered are the National Cancer Institute (NCI) selected 60 cell lines representing nine tumor types (breast, central nervous system, colon, leukemia, lung, melanoma, ovarian, prostate, and renal). In the above work the authors focus on gene expression response to stress. Gene expression data was derived by using two-colour spotted arrays (Qiagen library), consisting of 6,728 probes and 60 samples (GEO accession number GSE7505).

B. Data filtering

The most striking patterns in the radiation data were a set of genes up-regulated preferentially in the p53 wild-type lines, where p53 is a well-known tumour suppressor protein involved in prevention of cancer, and a set of cell cycle regulatory genes down-regulated across the entire NCI-60 panel. For that reason we estimated the differentially expressed genes which distinguish the p53 mutational samples from the wild type ones. Gene selection analysis was performed on the log2-normalized microarray data as found in GEO database. We applied the Significant Analysis of Microarrays (SAM) test to identify the differentially expressed genes (p-value < 0.05) given the p53 status - which is a major determinant of gene expression responses to ionizing radiation. A binary vector is constructed based on the significant p-values per gene for each test, namely 1 denotes significant genes and 0 otherwise.

In that way we can easily summarize the information of the statistical analysis conducted and we can easily correlate those results with gene ontology information derived by GO database. Then the set of genes was enriched using GO terms. The three categories of the GO hierarchy are biological processes, molecular functions, and sub-cellular locations. The resulting data set consists of 6,728 probes that are described by 1,260 GO terms. All statistical analysis was conducted in R (http://www.r-project.org/).

C. Results

To illustrate the interactive character of the SD we performed two iterations. In the first iteration the user specifies parameters such as a category of GO or the number of retrieved subgroups (Fig. 3). The service identifies the most interesting subgroups and displays them in the collaborative workbench (Fig. 4). Each retrieved subgroup is presented as a single element and can be discussed separately.

Table I shows the six most highly ranked estimated subgroups provided by the service given the over-expression of
genes considered. The different subgroups describe similar biological processes according to the GO database, and enable the user to test SD findings relative to other analysis outcomes, any information taken from other databases or from the literature available. In the second column of the table a Quality score is provided as a measure of each group’s statistical significance. Particularly, the quality of subgroup $S_i$ is computed as a ratio of differentially expressed genes that are associated with description terms over all genes associated with these terms. The results are further discussed in the following section by means of a collaboration instance paradigm showing everyday practices in bioscience field.

Within the Dicode workbench, the SD service offers an additional flexibility of incorporating feedback from other workbench users or/and filter the SD results by relying on the Quality measure provided. As shown in Fig. 4 different users can comment on the SD results of Table I, particularly the terms “positive regulation of apoptosis”, “multicellular organismal development” and “aerobic respiration” are marked as important biological procedures in the terms of cancer disease. Particularly, the first term has been reported to be very relevant for breast cancer disease, and it would be interesting to examine whether it is of high importance to multi-types cancer data examined here. For that reason it would be interesting to re-run the SD algorithm without including that particular term and compare the two outcomes. Other terms such as “protein binding” have a general meaning and are not thought to be explicitly relevant to the disease, thus there is scope in excluding them from future analysis. Users can launch other workbench services, for example the PubMed service, and decide on the relevance of the biological terms. When the user has decided upon which terms are biologically important for the analysis, he/she can re-run the algorithm whilst excluding different terms or subgroups, which then results in an altered prioritization of the remaining biologically important subgroups (Table II). In this case the terms “extracellular region”, “peptidase inhibitor activity”, “multicellular organismal development” are excluded causing a re-arrangement of SD’s results or even breaking down some subgroups to smaller groups. This procedure can be repeated multiple times and results from different runs can be compared.

V. COLLABORATION IN DICODE

Being fully integrated in the Dicode workbench, Collaboration and Decision Making Support services enable participants to argue and collaborate towards collectively considering various issues, the ultimate aim being to decide about which course of action has to be taken. Representative issues to be collaboratively solved include discussions about what data sources need to be investigated for the issue at hand or what policy should they follow after a particular interpretation of the outcomes of the data mining algorithms that are discussed in the previous section.

Collaboration in Dicode brings together two collaboration paradigms: the Web 2.0 paradigm, which builds on flexible rules favoring ease-of-use and human interpretable semantics, and the traditional decision support paradigm, which requires rigid rules that reduce ease-of-use but render machine interpretable semantics. To achieve this, our approach builds on a conceptual framework, where formality and the level of knowledge structuring during collaboration is not considered as a predefined and rigid property, but rather as an adaptable aspect that can be modified to meet the needs of the tasks at hand. By the term formality, we refer to the rules enforced by the system, with which all user actions must comply. Allowing formality to vary within the collaboration space, incremental formalization, i.e. a stepwise and controlled evolution from a mere collection of individual ideas and resources to the production of highly contextualized and interrelated knowledge artifacts and finally decisions, can be achieved [22]. Alternative Dicode “views” permit the abovementioned incremental formalization of collaboration. A view can be defined as a particular representation of the collaboration space, in which a consistent set of abstractions able to solve a particular organizational problem during collaboration is available. Our approach enables a rule-based transformation of
a view to another. Each view of a collaboration space provides
the necessary mechanisms to support a particular level of
formality. The more informal a view is, the more easiness-of-use is implied. At the same time, the actions that users
may perform are intuitive and not time consuming; however,
the overall context is human (and not system) interpretable.
On the other hand, the more formal a view is, easiness-of-use is reduced; actions permitted are less and less intuitive and
more time consuming. The overall context in this case is both
human and system interpretable. The views supported so far are:

- Discussion-forum view: A collaboration space is dis-
  played as a traditional web-based forum, where posts are
displayed in ascending chronological order. Users are able
to post new messages to the collaboration space, which
will appear at the end of the list of messages. The aim of
the discussion forum view is to allow the collection and
sharing of opinions without limiting the expressiveness
of participants.

- Mind-map view: A collaboration space is displayed as a
  mind map (Fig. 4), where users can interact with the items
uploaded so far. The map deploys a spatial metaphor
permitting the easy movement and arrangement of items
on the collaboration space. The aim of this view is to
support information triage [20], i.e. the process of sorting
and organizing through numerous relevant materials and
organizing them to meet the task at hand.

- Formal view: This view enables the posting of predefined
  knowledge items which adhere to a specific argumenta-
tion model (i.e., IBIS [17]). It invokes a set of dedicated
scoring and reasoning mechanisms aiming to aid users
conceive the outcome of a particular collaborative session
and receive support towards reaching a decision.

In the collaboration instance shown in Fig. 4, the collabo-
ration space is displayed in a “mind-map view”, where stake-
holders may organize their collaboration through dedicated
item types such as ideas, notes, comments and services. Ideas
stand for items that deserve further exploitation; they may
correspond to an alternative solution to the issue under consid-
eration and they usually trigger the evolution of the collabora-
tion. Notes are generally considered as items expressing ones
knowledge about the overall issue, an already asserted idea
or note. Comments are items that usually express less strong
statements and are uploaded to express some explanatory
text or point to some potentially useful information. Finally,
services are items that aim at integrating external services into
the workspace. Service items enable users to configure, start
and monitor the execution of the specified external services
from within the workspace and allow the automatic upload of
the results into the workspace when the execution of the
service terminates. Multimedia resources can also be uploaded
into the mind map view (the content of which can be displayed
upon request or can be directly embedded in the workspace).
The mind-map view deploys a spatial metaphor permitting the
easy movement and arrangement of items on the collaboration
space.

Fig. 4 shows an instance of a workspace, where three
researchers (with user names Jim, Alice and Neal) discuss
the results of the execution of the subgroup discovery service.
Initially, the three researchers create a new collaboration
workspace and decide to operate it in the mind-map view. Sub-
sequently, one of the researchers uploads into the collaboration
workspace an item of type “service” (top item of Fig. 4, “Sub-
group discovery service”) and configures it in order to allow
the execution of the subgroup discovery service. Configuration
of the “service” item includes specifying the URI for REST-
based services and all the necessary parameter (such as the
number of retrieved subgroups for SD). Once the service item
has been configured, the researchers can start its execution by
double clicking on the item. A visual cue on the item allows
the researchers to monitor the status of the execution (whether
it is still ongoing or has finished). When the execution of the
service terminates, the outcomes are automatically added into
the workspace. In Fig. 4, the 6 items connected to the service
item with arrows labeled “Result” correspond to the service
outcomes, i.e. the discovered subgroups (“inmate immune
response & positive regulation of apoptosis”, “extracellular
region & peptidase inhibitor activity & protein binding”, “aer-
obic respiration & cytoplasm & oxidation-reduction process”,
“cytosol & multicellular organismal development & nucleus &
perinuclear region of cytoplasm”, “DNA binding & epidermis
development”, “aerobic respiration & negative regulation of
transcription from RNA & cytoplasm & oxidation-reduction
process”) which are returned in XML. The researchers can
collaboratively discuss the obtained results by uploading items
that argue in favor (items connected with green arrows to the
result files in Fig. 4) or criticize the results (items connected
with red arrows to the result files in Fig. 4). Taking into
consideration the discussion of the results, the researchers
may decide to re-execute the subgroup discovery service with
different parameters such as omitting certain terms. At any
point in time, the researchers can change the view of the collab-
oration workspace and continue their collaboration in the
Discussion-forum or Formal view.

VI. EVALUATION OF THE DICODE SERVICES

Dicode services go through an ongoing evaluation process.
Their first evaluation round, which has been recently com-
pleted, aimed to assess a series of key success indicators
concerning the maturity of the technology used, as well as
the usability and acceptability of Dicode services in three
real-life contexts (clinico-genomic research, medical decision
making, and opinion mining from Web 2.0 data). Evaluators
were asked to read a service-specific scenario, experiment with
the Dicode services and fill in a mixed-type questionnaire
(responses expected were in both quantitative and qualitative
form). For the context related to this paper (clinico-genomic
research), the sample consisted of 58 evaluators with varying
background knowledge in bioscience fields. Answers to the
quantitative questions of the questionnaires were given in a
1 – 5 scale, where 1 stands for I strongly disagree and 5 for
I strongly agree. As far as the SD service is concerned, the evaluators agreed that the objectives of the service are met, that the service is novel to their knowledge and that they are satisfied with the performance of the service, although there were some concerns about the data intensive decision making issues (overall median and mode rating values equal to 3). Nevertheless evaluators seemed to agree on how well the data intensiveness issue was addressed by the Dicode’s Collaborative Decision Making Support services as a whole. From an analysis of qualitative evaluation results, it came out that a more user-friendly SD service interface accompanied by more detailed documentation of the service, will be an asset.

With respect to the overall quality of the Dicode workbench, the evaluators agreed (median: 4, mode: 4) that its objectives are met, that it is novel to their knowledge, that are satisfied with its performance and that they are overall satisfied with it. The evaluators were neutral (median: 3, mode: 3) with respect to whether the workbench addressed the data intensive decision making issues. As long as its acceptability is concerned, the evaluators agreed (median: 4, mode: 4) that the Dicode workbench has the full set of functions they expected, that its interface is pleasant and that they will recommend it to their peers/community.

VII. CONCLUSIONS

SD is one of the data mining services offered through a user-friendly widget in Dicode. As shown in the previous section, this service may meaningfully interoperate with the collaboration and decision making support services. Generally speaking, when one thinks about biomedical knowledge discovery based on data mining tools, it is important to set up a collaborative, interactive process, where users can easily decide about which data repositories should be considered, analyze the algorithmic results, discuss the weaknesses of the patterns that were identified, and set up a new iteration of the algorithm by defining other descriptive attributes or integrating other relevant data. The Dicode approach facilitates the above in various ways, such as:

- presenting the discovery patterns to the users in a form that allows their use as one piece of knowledge in the overall discussion process;
- making it easy for the users to give feedback to the algorithm, e.g. by allowing them to specify undesired attributes, non-interesting subgroups, or controlling the complexity of the output, such as the desirable length of the rules;
- making it easy in the overall system to select (and integrate new) data sets and attributes, both from external data sources as well as from the underlying discussion.

Taking into account the above evaluation feedback, we argue that our overall approach offers an innovative solution that reduces the overall complexity of real-life collaboration and decision making settings to a manageable level, thus permitting stakeholders to be more productive and concentrate on creative activities. A major future work direction concerns the improvement of Dicode services in terms of their documenta-
tion, user interfaces and performance. Another one concerns testing of these services in various data-intensive contexts towards further assessing their applicability and potential.

ACKNOWLEDGMENT

This publication has been produced in the context of the EU Collaborative Project DICODE - Mastering Data-Intensive Collaboration and Decision which is co-funded by the European Commission under the contract FP7-ICT-257184. This publication reflects only the authors views and the Community is not liable for any use that may be made of the information contained therein.

REFERENCES


