Relational and Semantic Data Mining for Biomedical Research

Nada Lavrač and Petra Kralj Novak
Jožef Stefan Institute and Jožef Stefan International Postgraduate School
Jamova 39, 1000 Ljubljana, Slovenia
Nada.Lavrač@ijs.si, http://kt.ijs.si/nada_lavrac/

Keywords: relational data mining, semantic data mining, biomedicine

Received: December 12, 2012

The paper presents a historical overview of data mining tools and applications in the field of biomedical research, developed at the Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia. It first outlines subgroup discovery and selected relational data mining approaches, with the emphasis on propositionalization and relational subgroup discovery, which prove to be effective for data analysis in biomedical applications. The core of this paper describes recently developed approaches to semantic data mining which enable the use of domain ontologies as background knowledge in data analysis. The use of the described tools is illustrated on selected biomedical applications.

Povzetek: Prispevek opisuje zgodovinski pregled razvoja orodij rudarjenja podatkov na področju biomedicine.

1 Introduction

Data analysis in biomedical applications aims at extracting potentially new relationships from data and providing insightful representations of detected relationships. Methods for symbolic data analysis are preferred since highly accurate but non-interpretable classifiers are frequently considered useless for medical practice. Subgroup discovery techniques [7, 20] are of interest to biomedical research, as they enable the discovery of patient subgroups from classified patient data, where the induced subgroup descriptions have the form of descriptive rules.

Let us illustrate the results of subgroup discovery in two biomedical applications. In the first application [4], the induced subgroup descriptions suggest how to select individuals for population screening, concerning high risk for coronary heart disease (CHD). One of the discovered rules describes a group of overweight female patients older than 63 years:

$$\begin{aligned}\text{High CHD Risk} \leftarrow\ & \text{gender = female } \&\\ & \text{age} > 63 \text{ years } \&\\ & \text{body mass index} > 25\mathrm{kg/m}^2\end{aligned}$$

In the second application [16], subgroup describing rules suggest genes that are characteristic for a given cancer type (leukemia), distinguishing it from other 13 cancer types (CNS, lung cancer, etc.):

$$\begin{aligned}\text{Leukemia} \leftarrow\ & \text{KIAA0128 } \textit{is diff\_expressed } \&\\ & \text{prostaglandin d2 synthase } \textit{is not diff\_expressed}\end{aligned}$$

The following sections presents the evolution of tools and techniques from inductive logic programming and relational data mining through special purpose systems for bioinformatics to general purpose semantic data mining approaches which enable the use of domain ontologies as background knowledge for data analysis. We conclude by describing new challenges in the focus of our current and future research.

2 Relational data mining for biomedical applications

We first present selected approaches to inductive logic programming (ILP) [11, 9] and relational data mining (RDM) [1] which showed a great potential for biomedical research due to their capacity of using background knowledge in the learning process. From the available background knowledge (encoded as logical facts or rules) and a set of classified examples (encoded as a set of logical facts), an ILP/RDM algorithm derives a hypothesized logic program which explains the positive examples. While ILP focuses on data and background knowledge represented in a logical formalism, RDM assumes that the background knowledge and data are encoded in a unique relational database format. Compared to standard data mining techniques where the input data is typically stored in a single data table (e.g., in Excel), the input to an ILP/RDM algorithm is thus much more complex.

Propositionalization [8] is a RDM approach, which has been applied in several biomedical applications. Consider relational subgroup discovery, an approach effectively implemented in the RSD algorithm [2]. RSD generates descriptive rules as conjunctions of terms which encode background knowledge concepts. RSD performs example-weighting [10] (used in the so-called weighted covering algorithm) and uses the weighted relative accuracy (WRAcc) measure as a heuristic for rule selection. For

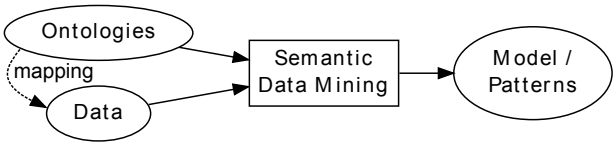

Figure 1: Semantic data mining schema

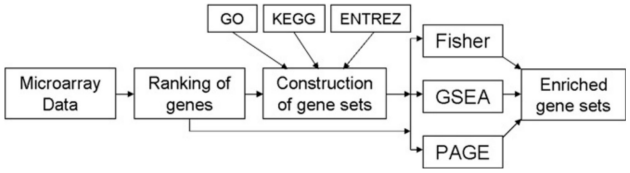

Figure 2: Schematic representation of SEGS.

example, an induced description of gene group A, discovered by RSD for the CNS (central nervous system) cancer class in the problem of distinguishing between 14 cancer types determines group A of differentially expressed genes in CNS as a conjunction of two relational features [17]: $geneGroup(A) \leftarrow f_i(A) \& f_k(A)$, where the two features, $f_i(A)$ and $f_k(A)$, constructed in the propositionalization step of RSD, are:

$f_i(A)$: *interaction(A,B)* &
process(B,'phosphorylation')
$f_k(A)$: *interaction(A,B)* &
process(B,'negative regulation of apoptosis') &
component(B,'intracellular membrane-bound organelle')

3 Semantic subgroup discovery

The RSD approach to relational subgroup discovery, which was successfully applied to mining microarray data [16], was the first step towards developing a novel data mining methodology, referred to as semantic subgroup discovery. The process of semantic data mining is illustrated in Figure 1.

The proposed semantic data mining methodology enables the generation of descriptive rules explaining the instances of a target class as conjunctions of ontology terms/concepts appearing in bioinformatics ontologies such as the well-known Gene Ontology (GO), KEGG and ENTREZ. An early approach to semantic subgroup discovery, named SEGS, is outlined below, followed by an outline of the SegMine methodology, which upgrades SEGS with a link discovery step.

3.1 Semantic subgroup discovery with SEGS

In many biomedical applications the goal of data analysis is gene set enrichment, i.e., finding groups of genes (gene sets) that are enriched, so that genes in the set are statistically significantly differentially expressed compared to the rest of the genes. Two well-known methods for testing the enrichment of gene sets include Gene Set Enrichment Analysis (GSEA, [15]) and Parametric Analysis of Gene Set Enrichment (PAGE, [6]). Originally, these methods use gene sets that are defined based on prior biological knowledge, e.g., published information about biochemical pathways, coexpression in previous experiments or Gene Ontology (GO) terms.

The RSD subgroup discovery approach combined with gene set enrichment analysis inspired the development of the SEGS algorithm (Searching for Enriched Gene Sets) [17], a specialized algorithm for semantic subgroup discovery for microarray data analysis. SEGS employs semantically annotated knowledge sources Gene Ontology (GO), the Kyoto Encyclopedia of Genes and Genomes (KEGG) and ENTREZ interactions, as background knowledge for semantic subgroup discovery. Based on this background knowledge, SEGS automatically formulates biological hypotheses: rules which define groups of differentially expressed genes. Finally, it estimates the relevance/significance of the formulated hypotheses on experimental microarray data. Compared to GSEA and PAGE, SEGS does not only test existing gene sets (defined by individual GO or KEGG terms), but constructs and tests also new gene sets, constructed by the combination of GO terms, KEGG terms, and also by taking into account the gene-gene interaction data from ENTREZ. The SEGS approach is outlined in Figure 2.

As it is infeasible to generate all the possible gene set descriptions in the given hypothesis language and evaluate each rule separately in the next step of the procedure, SEGS uses the topology of GO and KEGG to search the hypothesis space in a general-to-specific fashion to be able to reduce the search. Moreover, SEGS includes the ranking of genes (according to their differential expression based on the input microarray experiment) into the gene set generation phase (as shown in Figure 2) and counts the number of differentially expressed genes covered by each generated rule. If the number of covered differentially expressed genes is lower than a predefined threshold, the rule is eliminated and not specialized further, thus pruning large parts of the hypothesis space.

SEGS uses three statistical tests to evaluate the significance of the newly generated gene sets: Fisher's exact test, the GSEA method [15] and the PAGE method [6]. It then uses weights to combine the results of the three statistical tests.

Consider the application domain described in [14, 5], where data instances are gene expression profiles of patients belonging to two cancer classes, AML (acute myeloid leukemia) and ALL (acute lymphoblastic leukemia). Our goal is to uncover interesting patterns that can help to better understand the dependencies between the classes (cancer types) and the attributes (gene expressions values). The rules, shown in Figure 3, were generated from data on gene expression profiles obtained by the Affymetrix HU6800 microarray chip, containing probes for 6,817 genes, for 73 instances of AML or ALL class

Gene Set description	ES
Enriched in ALL	
1. ALL ← int(Func('zinc ion binding')& Comp('chromosomal part')& Proc('interphase of mitotic cell cycle'))	0.60
2. ALL ← Proc('DNA metabolism')	0.59
3. ALL ← int(Func('ATP binding')& Comp('chromosomal part')& Proc('DNA replication'))	0.55
Enriched in AML	
1. AML ← int(Func('metal ion binding')& Comp('cell surface')&	0.54
2. AML ← int(Comp('lysosome'))	0.53
3. AML ← Proc('inflammatory response')	0.51

Figure 3: Enriched gene set descriptions in the AML-ALL domain, together with their enrichment score (ES) [17].

labeled expression vectors. The rules are ranked according to the enrichment score, measuring the enrichment of differential expression of a set of genes, defined by the given conjunction of GO, KEGG and/or ENTREZ interactions.

3.2 SegMine: Combining SEGS and BioMine

The SegMine methodology [12], developed for exploratory analysis of microarray data, is performed through semantic subgroup discovery by SEGS, followed by link discovery and visualization by Biomine [3], an integrated annotated bioinformatics information resource of interlinked data. The SegMine methodology, illustrated in Figure 4, consists of gene ranking, hypothesis/rule generation by the SEGS method for enriched gene set construction, rule clustering, linking of the discovered gene sets to related biomedical databases for link discovery with Biomine, and Biomine sub-graph visualization.

The Biomine service is a valuable addition to SEGS, complementing our semantic subgroup discovery technology by additional explanatory potential due to additional Biomine graph visualization. Biomine is used through its web interface which allows for querying via Biomine named entities, such as a set of GO terms, resulting in a Biomine (sub)-graph, which can be visualized for exploratory purposes. A sample Biomine graph is shown in Figure 5, while the SegMine implementation in the Orange4WS workflow construction and execution platform [13] is shown in Figure 6. In [12], the utility of the SegMine methodology was demonstrated in two microarray data analysis applications: a well-known dataset from a clinical trial in acute lymphoblastic leukemia (ALL), and a dataset about the senescence in human mesenchymal stem cells (MSC). In the analysis of senescence in human stem cells, the use of SegMine resulted in three novel research hypotheses that can improve the understanding of the underlying mechanisms of senescence and identification of candidate marker genes.

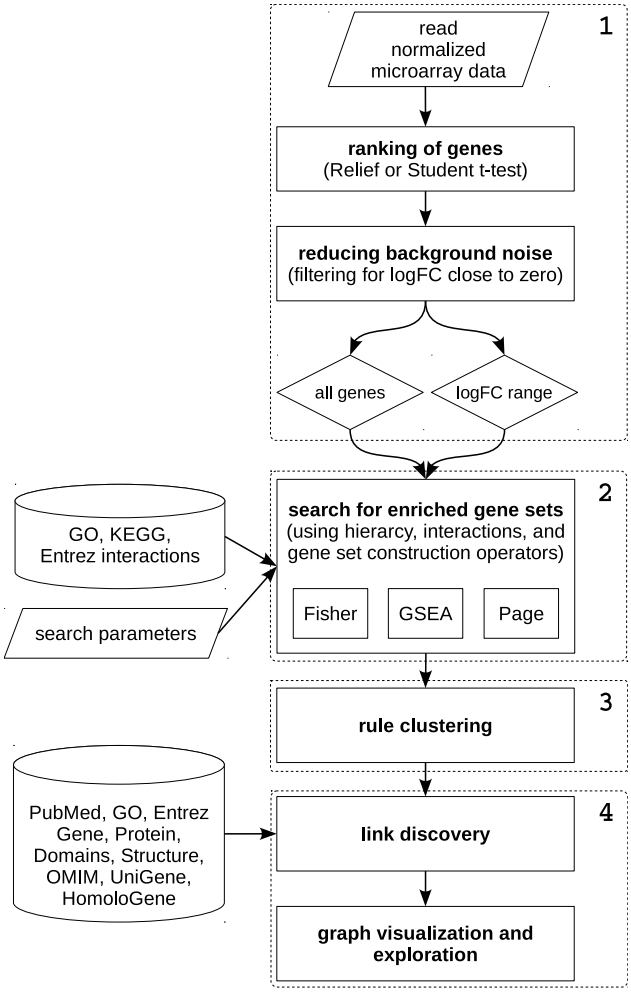

Figure 4: An overview of the SegMine methodology [12] emphasizing its four main steps: (1) data preprocessing, (2) search for differentially expressed gene sets, (3) clustering of rules describing differentially expressed gene sets, and (4) link discovery with graph visualization and exploration.

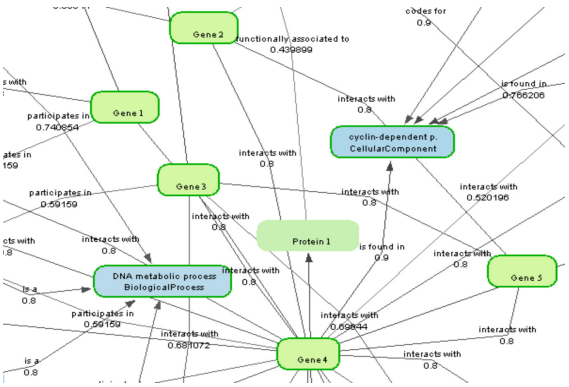

Figure 5: Biomine subgraph related to three genes from the enriched gene set constructed by SEGS.

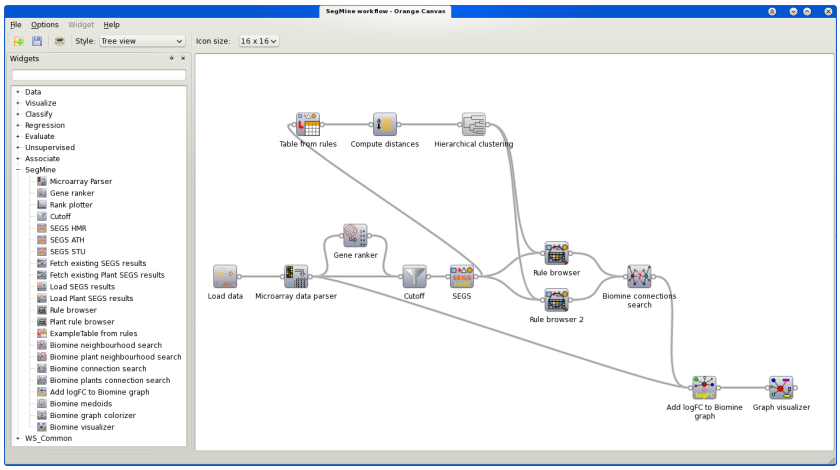

Figure 6: A screenshot of Orange4WS running a workflow of SegMine components [12].

4 General purpose semantic data mining

SEGS was the first special purpose semantic subgroup discovery algorithm developed. Recently, we developed two new general purpose semantic subgroup discovery systems: SDM-SEGS and SDM-Aleph [18]. SDM-SEGS is based on SEGS and can be used to discover subgroup descriptions from ranked data as well as from labeled data with the use of background knowledge in form of OWL ontologies. SDM-Aleph is based on the ILP system Aleph.[1] It was designed to be used in a similar way as SDM-SEGS. Unlike SDM-SEGS which is limited to four ontologies as input and only one additional *interacts* relationship, in SDM-Aleph any number of ontologies and additional relations between the input examples can be specified, which is due to the powerful underlying first-order logic formalism of the ILP system Aleph. SDM-SEGS and SDM-Aleph are implemented within a new semantic data mining toolkit, named SDM-Toolkit [18]. SDM-Toolkit has been made publicly available within the Orange4WS service-oriented data mining environment [13]. In [18], we illustrate the use of SDM-Toolkit tools for biomedical workflow construction and their execution in Orange4WS on the same two biomedical problem domains, ALL and hMSC, which were used in the evaluation of the utility of SegMine [12]. A qualitative evaluation of SDM-SEGS and SDM-Aleph, supported by experimental results and comparisons with SEGS, showed that SEGS and SDM-SEGS are more appropriate for data analysis in biomedical domains where rule specificity is desired, while SDM-Aleph is a more general purpose system, resulting in more general rules of lower precision.

Our recent work [19] also addresses semantic subgroup discovery, but focuses on a problem of explaining patient subgroups (e.g., similar patients, possibly all having a certain, yet unexplored cancer subtype) rather than explaining sets of differentially expressed genes characteristic for patients of a given class (cancer type) as a whole. This research is driven by a real-life problem of breast cancer patient analysis, motivated by the experts' assumption that there are several subtypes of breast cancer.

[1] http://www.cs.ox.ac.uk/activities/machinelearning/Aleph/

5 Conclusion

This paper presents a success story of three generations of data mining tools for biomedical research that use different forms of background knowledge. The paper presents the motivation and the evolution of ideas and techniques which were successfully applied in the field of biomedicine. A general-purpose semantic data mining toolkit is also presented, which offers numerous opportunities for applications where background knowledge in available in form of ontologies. All the presented tools are freely available online.

We envision further steps of development for semantic data mining. First, we foresee the usage of linked data as a general source of background knowledge used in semantic data mining. Second, we expect that the mining of knowledge encoded in ontologies will gain priority over mining the empirical data, which will, we believe, become a means of evaluation for the hypotheses generated from background knowledge.

Acknowledgement

We acknowledge numerous collaborators who have significantly contributed to this work: Dragan Gamberger, Filip Železny, Igor Trajkovski, Vid Podpečan, Igor Mozetič, Kristina Gruden, Hannu Toivonen and Anže Vavpetič.

The research presented in this paper was supported by the Slovenian Ministry of Higher Education, Science and Technology (grant no. P-103).

References

[1] S. Džeroski and N. Lavrač, editors. *Relational Data Mining*. Springer, New York, 2001.

[2] F. Železný and N. Lavrač. Propositionalization-based relational subgroup discovery with RSD. *Machine Learning*, 62(1-2):33–63, 2006.

[3] L. Eronen and H. Toivonen. Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC bioinformatics*, 13(1):119+, 2012.

[4] D. Gamberger and N. Lavrač. Expert-guided subgroup discovery: methodology and application. *J. Artif. Int. Res.*, 17(1):501–527, 2002.

[5] D. Gamberger, N. Lavrač, F. Železný, and J. Tolar. Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. *Journal of Biomedical Informatics*, 37(4):269–284, 2004.

[6] S.Y. Kim and D. J. Volsky. PAGE: Parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6:144, 2005.

[7] W. Klösgen. Explora: A multipattern and multistrategy discovery assistant. In *Advances in Knowledge Discovery and Data Mining*, pages 249–271. AAAI Press, Menlo Park, 1996.

[8] S. Kramer, N. Lavrač, and P. A. Flach. Propositionalization approaches to relational data mining. In N. Lavrač and S. Džeroski, editors, *Relational Data Mining*, pages 262–286. Springer, 2001.

[9] N. Lavrač and S. Džeroski. *Inductive Logic Programming: Techniques and Applications*. Ellis Horwood, New York, 1994.

[10] N. Lavrač, B. Kavšek, P. Flach, L. Todorovski, and S. Wrobel. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.

[11] S. Muggleton, editor. *Inductive Logic Programming*. Academic Press, London, 1992.

[12] V. Podpečan, N. Lavrač, I. Mozetič, P. Kralj Novak, I. Trajkovski, L. Langohr, K. Kulovesi, H. Toivonen, M. Petek, H. Motaln, and K. Gruden. SegMine workflows for semantic microarray data analysis in Orange4WS. *BMC Bioinformatics*, 12:416, 2011.

[13] V. Podpečan, M. Zemenova, and N. Lavrač. Orange4WS environment for service-oriented data mining. *Comput. J.*, 55(1):82–98, 2012.

[14] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*, 98(26):15149–15154, 2001.

[15] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, 2005.

[16] I. Trajkovski, F. Železný, N. Lavrač, and J. Tolar. Learning relational descriptions of differentially expressed gene groups. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 38(1):16–25, 2008.

[17] I. Trajkovski, N. Lavrač, and J. Tolar. SEGS: Search for enriched gene sets in microarray data. *Journal of Biomedical Informatics*, 41(4):588–601, 2008.

[18] A. Vavpetič and N. Lavrač. Semantic subgroup discovery systems and workflows in the SDM-Toolkit. *The Computer Journal*, 2012.

[19] A. Vavpetič, V. Podpečan, S. Meganck, and N. Lavrač. Explaining subgroups through ontologies. In *PRICAI2012: Proceedings of the National Academy of Science*, volume 7458, pages 625–636, 2012.

[20] S. Wrobel. An algorithm for multi-relational discovery of subgroups. In *Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery*, PKDD '97, pages 78–87, London, UK, UK, 1997. Springer-Verlag.