Relational and Semantic Data Mining for Biomedical Research

Nada Lavrač and Petra Kralj Novak
Jožef Stefan Institute and International Postgraduate School Jožef Stefan
Jamova 39, 1000 Ljubljana, Slovenia
e-mail: nada.lavrac@ijs.si

Abstract

The paper 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 described techniques and tools is illustrated on selected biomedical applications.

1. Introduction

In data mining and knowledge discovery [1], there are usually large amounts of data available and the data mining expert is confronted with a task of constructing a hypothesis – a predictive model or a set of descriptive patterns – induced from the data. The following steps of the model/patterns construction are usually performed: *Empirical Data* → *Preprocessing* → *Preprocessed Data* → *Data Mining* → *Hypothesis.* Results evaluation and interpretation by the expert can lead to iterative refinements of induced model/patterns.

The starting point for this paper are *subgroup discovery techniques* [2][3]. Subgroup discovery techniques 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 subgroups in two biomedical applications.

In the first application [4], the induced subgroup descriptions suggest the general practitioner 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:

High CHD Risk ← *gender = female &*
age > 63 years & body mass index > 25 kg/m^2

In the second application [5], 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.):

Leukemia ← *KIAA0128 is diff_expressed &*
prostaglandin d2 synthase is not diff_expressed

We continue by presenting selected approaches to *inductive logic programming* (ILP) [6][7] and *relational data mining* (RDM) [8] which also have 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 exemplified above, 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 [9] is a RDM approach, which has been applied in several biomedical applications. Consider *relational subgroup discovery,* an approach effectively implemented in the RSD algorithm [10]. RSD generates descriptive rules as conjunctions of terms which encode background knowledge concepts. For instance, 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 [11]: *geneGroup(A)* ← $f_i(A)$ & $f_k(A)$, 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'),* and $f_k(A)$: *interaction(A,B) & process(B,'negative regulation of apoptosis') & component(B,'intracellular membrane-bound organelle').*

The next section presents and overview of recently developed approaches to semantic data mining which enable the use of domain ontologies as background knowledge for data analysis, where the use of described techniques is illustrated on a biomedical application. We conclude by describing new challenges in the focus of our current and future research.

3. Semantic subgroup discovery

The presented approach to relational subgroup discovery, which has successfully used RSD to mine bioinformatics data [11], was the first step towards developing a novel data mining methodology, referred to as *semantic subgroup discovery*. The main steps in semantic pattern construction are the following: Empirical Data and Ontologies → *Preprocessing* → Preprocessed Data and Ontological Concepts → *Semantic Subgroup Discovery* → Semantic Descriptions Explaining the Discovered Subgroups.

This proposed 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 the outline of the recently developed SegMine methodology.

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, [13]) and *Parametric Analysis of Gene Set Enrichment* (PAGE, [14]). Originally, these methods take terms (gene sets) from the Gene Ontology (GO), the Kyoto Encyclopedia of Genes and Genomes (KEGG) and ENTREZ interactions, and test whether the genes annotated by a specific term are statistically significantly differentially expressed in the given dataset.

The RSD semantic subgroup discovery approach was adapted to gene set enrichment analysis in the SEGS algorithm (*Searching for Enriched Gene Sets*) [12], a specialized algorithm for semantic subgroup discovery for microarray data analysis.

SEGS employs semantically annotated knowledge sources GO, KEGG and ENTREZ, 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 1.

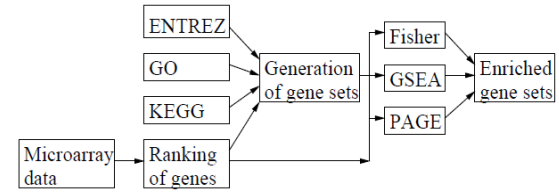

Figure 1. Schematic representation of SEGS.

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 1) 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.

For rule evaluation, SEGS uses three statistical tests to test the significance of the newly generated gene sets: Fisher's exact test, the GSEA method [13] and the PAGE method [14]. It then uses weights to combine the results of the three statistical tests.

Consider the application domain described in [15] and [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 Table 1, 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 labeled expression vectors. The rules in Table 1 are ranked

according to the enrichment score (ES) measuring the enrichment of differential expression of a set of genes, annotated by the given conjunction of GO, KEGG and/or ENTREZ terms.

Table 1. Enriched gene set descriptions in AML-ALL, together with their enrichment score (ES) [12].

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

3.2. SegMine: Combining SEGS and BioMine

The SegMine methodology [16], developed for exploratory analysis of microarray data, is performed through semantic subgroup discovery by SEGS, followed by link discovery and visualization by Biomine [17], an integrated annotated bioinformatics information resource of interlinked data. The SegMine methodology, illustrated in Figure 2, consists of the gene ranking, hypothesis/rule generation by the SEGS method for enriched gene set construction, linking of the discovered gene sets to related biomedical databases for link discovery with Biomine, and Biomine sub-graph visualization.

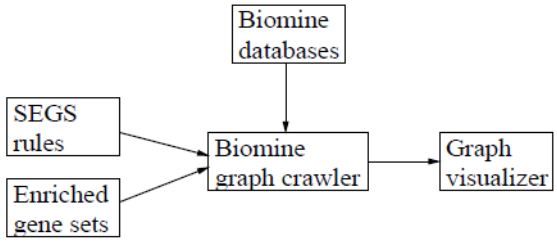

Figure 2. SegMine wokflow (focusing on Biomine, as details of the SEGS workflow are shown in Figure 1).

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 3, while the SegMine implementation in the Orange4WS workflow construction and execution platform [18] is shown in Figure 4. In [16], 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.

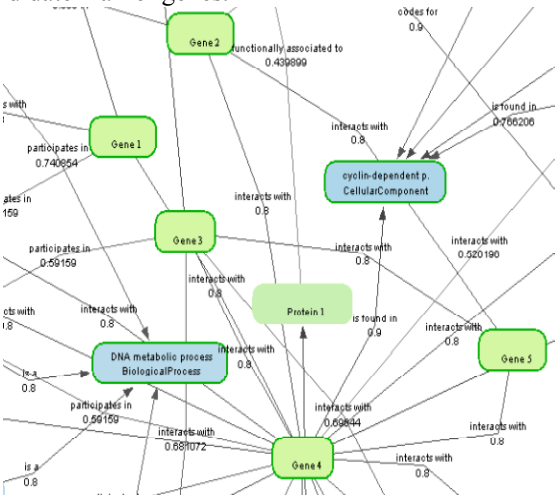

Figure 3. Biomine subgraph related to three genes from the enriched gene set constructed by SEGS.

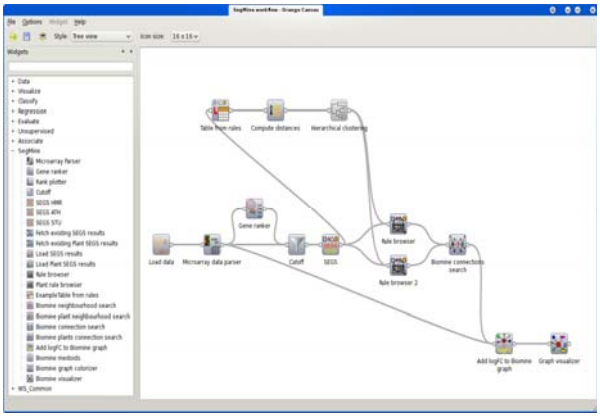

Figure 4. A screenshot of Orange4WS running a workflow of SegMine components [16].

4. Conclusions

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, and implemented them within a new semantic data mining toolkit, named SDM-Toolkit [19]. SDM-Toolkit has been made publicly available within the Orange4WS service-oriented data mining environment [18].

In [19], 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 [16]. 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 higher precision.

Our recent work [20] 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 motivated by a real-life problem of breast cancer patient analysis, motivated by the experts' assumption that there are several subtypes of breast cancer.

Acknowledgments

To conclude, 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.

References

[1] U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery: An Overview", in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.) Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, 1996, pp. 1–34.

[2] W. Kloesgen, "EXPLORA: A Multipattern and Multistrategy Discovery Assistant", in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.) Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, 1996, pp. 249–271.

[3] S. Wrobel, "An Algorithm for Multi-relational Discovery of Subgroups", in J. Komorowski and J. Zytkow (Eds.) Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-97), Springer, 1997, pp. 78–87.

[4] D. Gamberger and N. Lavrač, "Expert-Guided Subgroup Discovery: Methodology and Application". Journal of Artificial Intelligence Research, 2002, 17, pp. 501–527.

[5] D. Gamberger, N. Lavrač, F. Železny, and J. Tolar, "Induction of Comprehensible Models for Gene Expression Datasets by the Subgroup Discovery Methodology", Journal of Biomedical Informatics, 2004, 37, pp. 269–284.

[6] S. Muggleton (Ed.), Inductive Logic Programming, Academic Press, London, 1992.

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

[8] S. Džeroski and N. Lavrač (Eds.) Relational Data Mining, Springer, 2001.

[9] S. Kramer, N. Lavrač and P. Flach, "Propositionalization Approaches to Relational Data Mining", in S. Džeroski and N. Lavrač (Eds.) Relational Data Mining, Springer, 2001, 262-286.

[10] F. Železny and N. Lavrač. "Propositionalization-based Relational Subgroup Discovery with RSD", Machine Learning, 2007, 62(1–2), pp. 33–63.

[11] I. Trajkovski, F. Železny, N. Lavrač, and J. Tolar. "Learning Relational descriptions of Differentially Expressed Gene Groups", IEEE Transactions of Systems, Man and Cybernetics C, Special issue on Intelligent Computation for Bioinformatics, 2008a, 38(1), pp. 16–25.

[12] I. Trajkovski, N. Lavrač, and J. Tolar, "SEGS: Search for Enriched Gene Sets in Microarray Data, Journal of Biomedical Informatics, 2008b, 41(4), pp. 588–601.

[13] P. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, et al., "Gene set enrichment analysis: A knowledge based approach for interpreting genome-wide expression profiles', in Proceedings of the National Academy of Science, USA, 2005, 102(43), pp. 15545–15550.

[14] S.Y. Kim and D.J. Volsky, "PAGE: Parametric Analysis of Gene Set Enrichment", BMC Bioinformatics, 2005, 6:144.

[15] S. Ramaswamy et al. "Multiclass cancer diagnosis using tumor gene expression signitures", in Proceedings of the National Academy of Science, USA, 2001, 98(26), pp. 15149–15154.

[16] V. Podpečan, N. Lavrač, I. Mozetič et al., "SegMine workflows for semantic microarray data analysis in Orange4WS", BMC Bioinformatics 2011, 12(416).

[17] L. Eronen and H. Toivonen, "Biomine: predicting links between biological entities using network models of heterogeneous databases", BMC Bioinformatics, 2012, 13 (119).

[18] V. Podpečan, M. Zemenova, and N. Lavrač, "Orange4WS Environment for Service-Oriented Data Mining", The Computer Journal, 2012, 55, pp. 82-98.

[19] A. Vavpetič and N. Lavrač, "Semantic Subgroup Discovery Systems and Workflows in the SDM-Toolkit", The Computer Journal, 2012; doi: 10.1093/comjnl/bxs057

[20] A. Vavpetič, V. Podpečan, S. Meganck and N. Lavrač, "Explaining subgroups through ontologies", in Proceedings of The Pacific-Rim Conference on Artificial Intelligence, PRICAI-2012, 2012.