Facilitating Clinico-Genomic Knowledge Discovery by Automatic Selection of KDD Processes

Natalja Punko NATALJA.PUNKO@IAIS.FRAUNHOFER.DE
Stefan Rüping STEFAN.RUEPING@IAIS.FRAUNHOFER.DE
Stefan Wrobel STEFAN.WROBEL@IAIS.FRAUNHOFER.DE

Keywords: Clinico-Genomic Knowledge Discovery, Meta Learning, Similarity Learning, Ontologies

Fraunhofer IAIS, Schloss Birlinghoven, 53754 St. Augustin, Germany

Abstract

The analysis of clinico-genomic data poses complex problems for machine learning. As high volumes of data can be generated easily, selecting the most suitable KDD-process for the problem at hand becomes increasingly hard, even for experienced researchers. The main idea of this paper is to facilitate process selection by representing each data set by a graph based on the ontology that describes data set attributes, and to apply graph mining methods to perform a similarity search. Some new measures for an effective comparison of a data set graph induced from the ontology are proposed. The effectiveness of the proposed approach is evaluated on three datasets. The results show that using ontology-based characteristics leads to improving the characterization of a data set.

1. Introduction

In recent years, developments in distributed architectures, such as Grid technologies, have led to a vast increase in computing power and storage space that is available to end-users. The drawback of easily generating large volumes of data is that its analysis becomes increasingly hard and that even experienced researchers are facing problems to keep track of the state-of-the-art for a given analysis problem. Instead of being supported by information technology, the user is overwhelmed by an enormous volume of information. Hence, the user needs assistance in navigating the space of appropriate KDD processes. The knowledge about solutions that were successfully used in the past for a similar problem provides a good base for building a supportive system for the user.

This work aims to develop an approach that assists the user without specific data mining experience to select a suitable KDD process. The fundamental idea that underlies this work is that data sets similar in content are likely to be analyzed with the same KDD processes.

This assumption allows to reduce the problem of KDD process selection to searching for similar data sets. However, measuring the similarity between data sets is not a simple task. What is needed is a description of a data set that allows efficient comparison. Each data set can be represented by a tuple D =(A,T) consisting of attributes (A) and data (T). The current approach to describe a data set used in meta-learning rely on a set of characteristics that can be derived directly from data (Peng et al., 2002; Kalousis & Hilario, Januar 2001; METAL, 2002). There is still a need to improve the characteristics by developing more informative ones. This work aims to extend the data set description by involving new characteristics which focus on attributes to include semantic information. We propose to describe a data set by using the relationship between attributes captured in an ontology. For each new data set based on the developed characteristics we can find the most similar existing data set with known good analysis process. The top k of these processes can be applied to the new data set. The next task is the method for comparison two dataset based on the developed characteristics.

The main results of this work are the development of ontology-based characteristics to improve the dataset description, the development of a method allowing effective comparison of datasets based on the developed

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

characteristics, and the evaluation of the approach.

This paper is organized as follows: The next section gives a short introduction to a real-life Grid project, on which this work is based. Section 3 presents a brief overview of previous research. The novel approach for characterizing and comparing datasets is presented in Section 4. Experiments and results illustrating the effectiveness of new the approach are presented in Section 6. Section 7 concludes the paper.

2. The ACGT Project

In this section, we give a short introduction to the ACGT[1] project, as we feel that it is necessary to understand some of the implications of eScience applications to understand our approach.

In recent years, the rapid development of high throughput genomics and post-genomics technologies has provided clinicians fighting cancer with new discovery paths and has opened the possibility to develop patient-specific treatment strategies.

However, the amount of information now available for each patient (e.g. in microarray context from 10,000s to 100,000s of variables summarizing up to millions of array features) has rendered difficult the isolation of the clinically relevant information from all available data. Considering the current size of clinical trials (hundreds of patients), there is a clear need, both from the viewpoint of the fundamental research and from that of the treatment of individual patients, for a data analysis environment that allows the exploitation of this enormous pool of data (Wegener et al., 2007).

Advancing Clinico-Genomics Trials on Cancer (ACGT) project aims at developing an open-source IT infrastructure to provide the biomedical community with the tools needed to integrate complex clinical information and make a concrete step towards the tailorization of treatment to the patient.

The ACGT architecture can - for the purposes of this paper - be described as a workflow enacting environment for based on distributed computing and Grid technologies. It will provide a large set of data analysis operators from which suitable analytic workflows can be constructed. Grid technology allows to securely execute computational intensive, geographically distributed, parallel workflows, which forms an excellent basis for conducting knowledge discovery experiments. Note that while it is infeasible to explore the complete space of possible workflows, it is still very easy to execute and compare a larger set (say, 20) of workflows in parallel. This simplifies the problem of KDD process selection somewhat, as we are no longer required to reliably identify the optimal one. Instead, it suffices to identify a good set of candidate workflows, such that the optimal one is among the top k. This motivates our approach of viewing the process selection problem essentially as a top k ranking problem.

[1] http://www.eu-acgt.org

2.1. The ACGT Master Ontology on Cancer

Data access within ACGT is based on a semantic Grid infrastructure to enable integrated access to multilevel biomedical data (Tsiknakis et al., 2008). The basis of this approach is the ACGT Master Ontology on Cancer, which can be used to describe and query stored data sets (Brochhausen et al., June 17 19 2008). The Master Ontology consisting of 1109 classes and 121 restrictions. An overview can be seen in Figure 1.

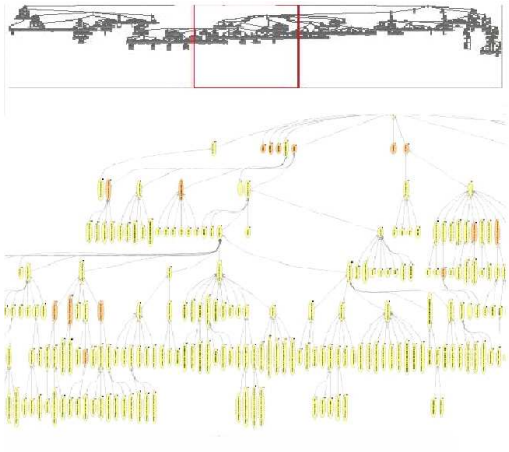

Figure 1. Excerpt from the ACGT Master Ontology

Note that while strict legal and ethical guidelines on privacy severely restrict the storage and processing of actual patient data except for clearly defined purposes, much less restrictions exists for meta data descriptions of these data sets and statistics about KDD process executions. For this reason, we base our algorithm on an analytic database of ontologic data descriptions and process performance measures.

3. Related Work

Several approaches to describe a dataset aim to assist a user by selecting an appropriate algorithm. There exist algorithms that perform good on some datasets but can't be applied on the others.

Extensive research to develop characteristics describing a dataset has been performed in the context of StatLog Project (Kalousis & Hilario, 2001). Descrip-

tion was created using a number of simple, statistical and information-theoretical measures estimated directly from a dataset. Developed characteristics have been repeatedly used for algorithm predictions in many works (Brazdil et al., 1994; Sohn, 1999; Todorvoski & Dzeroski, 1999).

(Engels & Theusinger, 1998) presents a tool for automatic computation of dataset characteristics developed in StatLog Project called DCT - Data Characterization Tool. In context of METAL Project (Köpf et al., 2000) this set of measures has been extended by some new characteristics. The current version of DCT is able to compute about 50 basic measures for each dataset, and additional measures for each attribute. The METAL "Meta-Learning Assistant" available under (Köpf et al., 2000) is a web-enabled prototype providing support for the user of machine learning tools.

3.1. Meta Learning

In this section we shortly present the meta-data-based approach to dataset characterization that was used as a part of the experiments in this paper. According to the idea that the characteristics of a dataset provide some important information for an algorithm selection two basic tasks of meta-learning can be defined: developing a dataset description and deriving knowledge about correlations between particular characteristics and the performance of several algorithms.

Some basic strategies to describe the data set are presented below. Here, we have analyzed in detail the characteristics developed in the context of METAL Project. They are three basic sets of measures that can be estimated directly from the data set. Recall that the data set consists of a set of attributes and data D=(A,T). Characteristics that were constructed include simple characteristics (number of classes, frequency symbolic and numeric attributes, distribution of classes, accuracy of the default-class), statistical characteristics (number of significant discriminant functions, Willk's Lambda), and information-theoretical characteristics (entropy of classes, entropy of attributes, joint Entropy, equivalent number of attributes, mutual information, noise).

These characteristics are computed using only one aspect of a dataset, namely the data. As mentioned above, we propose to concentrate on the semantic information contained in the attributes of the dataset. The semantic knowledge in the form of ontologies provides a powerful support for the techniques used for managing data in recent years. However, there might be attributes in the data set that syntactically are different from one another but semantically they are equivalent and express the same concepts. We use an ontology as a semantic layer to describe the semantic relationships between the dataset attributes in order to extract additional information about the dataset.

To use ontologies in our approach the following requirements should be met:

1. An ontology should be broad enough to describe a wide range of concepts belonging to the subject area.
2. An ontology should be specific enough to describe correctly the relationships between the attributes of the dataset belonging to this subject area.

As we can see the ACGT Master Ontology describing data sets demonstrates a realistic situation where both of these requirements are met.

3.2. Graph Mining

Modeling complex data with the help of graphs has become an active research area in the last few years. Graph models are successfully used in a broad range of applications, such as web analysis, drug discovery and compound synthesis. While data mining is concerned with frequent data values, graph mining deals with frequent subgraphs and common specific topologies. Traditional approaches to characterize the graph to discover typical patterns are realized in two different ways, namely by structural features (Nakano et al., 2007) and by the number of occurrences of certain substructures (e.g. in drug discovery (Deshpande et al., 2003)). The first approach is used in our work to represent a dataset modeled as a graph by a set of its topological properties.

4. New Approach

4.1. Representing datasets as graphs

Based on the ontology describing a domain of dataset attributes we can transform a dataset to a graph. Basic steps are as follows.

4.1.1. Construction of vertices and edges

As first step the dataset attributes are mapped to the corresponding concepts in the ontology. Formally, an ontology $\mathcal{O} = (C, \leq_C, \sigma, R, \leq_R)$ consists of a set C of concepts that correspond to classes, a set R of relations among class members, a partial order $\leq_C$ on C (called taxonomy), a partial order $\leq_R$ on R (called relations hierarchy), and a function $\sigma : R \rightarrow C \times C$ that maps the relations to the concepts type. For a dataset

$D = (A, T)$ we define the dataset graph as $G = (V, E)$, where $V = A$ and $E = \{(v, v')|(v, v') \in R, v, v' \in V\}$.

The dataset attributes can be seen as a set of nodes. Each node is labeled with the corresponding attribute name. The set of edges can be modeled by the relationships between corresponding concepts in the ontology. Two vertices in the dataset graph become adjacent if there exists a direct link between the related concepts in the ontology. In order to include in a dataset graph the information about the ontology structure each edge is associated with a weight value. We define the weight of the edge (v, v') as its depth in a taxonomy that is included in the ontology. A visualization of this step is shown in Figure 2.

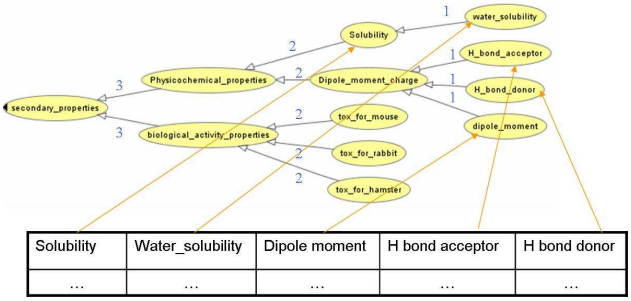

Figure 2. Mapping attributes to ontology concepts.

In the case that no direct relationships between the attributes in the ontology are available the dataset graph consists only of the vertices, while the set of edges is empty. Such a structure is not suited for the comparison. To remedy this problem we perform a primary graph extension to produce a connected graph. A graph is called connected if there is a path from any node to any other node in the graph.

4.1.2. Extension of the nodes set

The node set is extended with the following nodes:

$$V^* = \{v^*|v^* \in shPath(v, v'), v, v' \in V\} \qquad (1)$$

where $shPath(v, v')$ is the shortest path between v and v'. V^* consists of the vertices lying on the shortest path between every two vertices that are not directly adjacent in the ontology. The primary set of nodes is extended with the "missing" concepts that connect the attributes in the ontology.

4.1.3. Extension of the primary graph

As the last step the set of edges is extended with the edges that join the vertices from V and V^* in the ontology graph. Figure 3 gives an example of an extended graph.

$$E^* = \{(v', v'')|v', v'' \in V \cup V^*\} \qquad (2)$$

The final dataset graph G_A is defined as follows:

$$G_A = (V \cup V^*, E \cup E^*) \qquad (3)$$

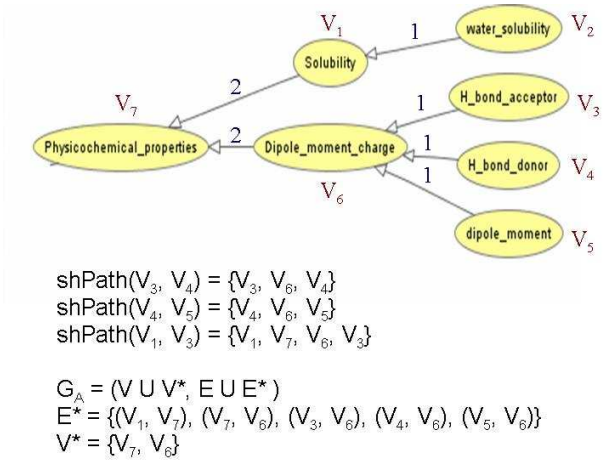

Figure 3. Extended data set graph.

4.2. Local Measures

As described above the dataset is represented by a graph. The connections of the graph vertices corresponds to the structural information of the dataset. Based on this concept we characterize the dataset by the following topological properties of the induced graph: characteristic path length (average length of the shortest path between two vertices in a graph), edges distribution (average number of edges connected to the vertices), diameter (maximal distance between two vertices in the graph), distance to the target attribute (minimal distance between the target attribute and the other verteces), and connectivity (also called the Beta-Index) The local similarity measures are defined as the linear difference of the corresponding graph properties.

4.3. Distance measures for dataset graphs

Next we introduce some new measures computed by directly comparing of two dataset graphs. The following distance measures for two dataset graphs are proposed:

Relative size of maximal common subgraph: This measure estimates common attributes and relationships.

$$dist_1(G_1, G_2) = 1 - \frac{|MGS(G_1, G_2)|}{max(|G_1|, |G_2|)} \qquad (4)$$

Relative sum of common subgraphs: Besides the maximal common subgraph all common subgraphs provide useful information regarding graphs similarity. At first we find the set of all common subgraphs for G_1 and G_2:

$$SUB(G_1, G_2) = \{S_1, S_2, \ldots, S_n\},\ where\ S_i \cap S_j = \emptyset \quad (5)$$

The subgraph size is divided by the maximal graph size to normalize the distance value. The final distance is computed as follows:

$$dist_2(G_1, G_2) = 1 - \sum_{i=1, S_i \in SUB}^{n} \frac{|S_i|}{max(|G_1|, |G_2|)} \quad (6)$$

4.4. Ontology-based measures

Based on the ontology we can examine the content similarity of dataset attributes. In the case that attributes with different names correspond to the same ontology concept, they are equal. If there are many equal attributes, then datasets are likely to be very similar. We define the distance measures based on the distance between two concepts in the ontology.

To determine the distance between target attributes we use the normalised weighted path length between the corresponding nodes in the ontology. It can be computed by adding the weights of the corresponding edges.

$$dist_3(c_1, c_2) = \frac{length_w(c_1, c_2)}{D} \quad (7)$$

where $dist_3(c_1, c_2)$ is the distance between the target attributes c_1 and c_2, $length_w(c_1, c_2)$ is the weighted length of the shortest path between c_1 and c_2, and D is the maximal weighted path length in the ontology.

For example, based on the ontology given by Figure 2 we compute distance between attributes "water_solubility" (c_1) and "dipole_moment" (c_2) as follows:

$$dist_3(c_1, c_2) = \frac{(1+2+2+1)}{(1+2+3+3+2)} \quad (8)$$

To measure the average similarity of dataset attributes, attributes are compared pairwise using (7) and the corresponding values are averaged.

4.5. Distance between sets of attributes

In order to create a connected graph the primary set of attributes is extended by some vertices. Therefore, primary and extended attribute sets are different. We introduce two additional distance measures respecting this problem by comparison of primary attribute sets, the Jaccard Distance and the Overlap-coefficient. Despite the name, the Jaccard distance is a similarity measure and estimates the ratio of common attributes relative to all attributes:

$$d_{Jaccard}(A_1, A_2) = 1 - sim(A_1, A_2) = 1 - \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|} \quad (9)$$

The Overlap-coefficient estimates the ratio of common attributes relative to the maximal attributes set.

$$d_{Overlap}(A_1, A_2) = 1 - sim(A_1, A_2) = 1 - \frac{|A_1 \cap A_2|}{max(|A_1|, |A_2|)} \quad (10)$$

4.6. Global Similarity Measure

The global similarity of datasets (σ^*) is implemented as a weighted sum of the base similarity measures (d_i) from Sections 4.2, 4.3, 4.4 and 4.5.

$$\sigma^* = 1 - dist^* = 1 - (d_1 \cdot w_1 + d_2 \cdot w_2 + \cdots + d_k \cdot w_k) \quad (11)$$

Weights (w_i) can be found using the supervised learning algorithm described below. The motivation of our approach is based on the assumption that similar data sets are likely to have similar rankings of KDD processes. In order to define a learning problem, we turn this assumption around and actually define data sets to be similar when they have the same ranking.

Definition: Given a set of KDD processes $\mathcal{P} = (P_i)_{i=1}^n$, consider the set $\mathcal{D}$ of all data sets on which all P_i can be applied. Let q be some real-valued quality measure on the output $P(D)$ of a process P on a data set $D \in \mathcal{D}$. Two datasets D and D' are called similar, if the ranking of the qualities of the processes in $\mathcal{P}$ on D and D' is similar. More specific, let the similarity measure $\sigma_{true}(D, D')$ be defined by the correlation of the rankings given by the processes in $\mathcal{P}$.

With this definition, we have reduced a problem of selecting the most suitable KDD process for a given data set to the search of the most similar dataset. A simple, but computationally too expensive solution would be to simply calculate σ_{true}. In order to find practically feasible solution, we now address the learning task of finding a similarity measure σ^* that is defined on features of the data set D only, i.e. that can be computed without the execution of a KDD process, such that σ^* mimics σ_{true} as closely as possible. The visualization of the learning process is shown in Figure 4. To measure the agreement between rankings of two datasets, denoted as σ_{true} we use Spearman's rank correlation coefficient (Neave & Worthington, 1992).

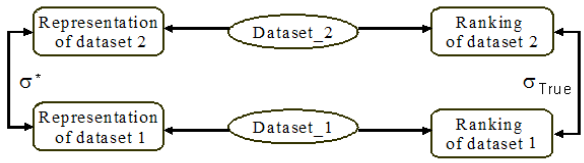

Figure 4. Data sets similarity learning.

Algorithm 1 Distance Measure Learning

Input: similar S and dissimilar D pairs.
Initialise α with random values.
Compute $g(A)$
Minimize $g(A)$ using Newton-Raphson method.
Output: α - weights vector.

4.7. Learning algorithm

In this work we used a learning algorithm proposed by (Xing et al., 2003). For the similarity measure a few data sets with known σ_{true} were used to define the sets of truly similar and dissimilar datasets. Ordered by σ_{true} the top k data set pairs are defined as the set of similar example sets, while bottom k data set pairs form the dissimilar example set.

The algorithm can be described as follows: For the vector representation of the data we can define the distance metric as:

$$d(x,y) = d_A(x,y) = ||x-y||_A = \sqrt{(x-y)^T A(x-y)}, \quad (12)$$

where A is positive semi-definite matrix, in other words, a symmetric matrix with non negative eigenvalues. The goal is to minimize the distance between similar pairs.

$$min_A \sum_{(x_i,x_j)\in\mathcal{S}} ||x_i - x_j||_A^2. \quad (13)$$

with the following constraint:

$$\sum_{(x_i,x_j)\in\mathcal{D}} ||x_i - x_j||_A \geq 1 \quad (14)$$

where S and D are the sets of similar and dissimilar pairs respectively. We apply the case to learn a diagonal maxtrix $A = diag(A_{11}, A_{22}, \ldots, A_{nn})$. This problem is solved using Newton-Raphson method. The summary of our approaches is given in Algorithm 1.

As a reference approach we directly use SVMs (Cortes & Vapnik, Sept 1995) on the classification problem defined by S and D.

5. Experiments

In this section we evaluate the effectiveness of the proposed characteristics. Before the start of the evaluation we have to determine the experimental setup. First, a set of data sets is needed. Second, computing developed characteristics from chosen data sets assumes that the ontology describing the attribute sets is available. Third, we need a knowledge database containing multiple data sets together with the quality measure of KDD processes that were applied to these data sets. In order to facilitate the evaluation of the proposed approach we limit the KDD processes to a single classification algorithm. Note that as our method does not use any properties of the process itself, the application to more complex KDD processes is straight-forward.

5.1. Data Sets

Finding appropriate data sets for the evaluation of our approach is a hard task, as the success of our approach depends on the availability of a set of connected data sets and a high-quality ontology to provide the necessary semantics. On the one hand, standard data sets usually do not come with ontologies, on the other hand, in the domain where a high-quality ontology was available to us (the ACGT ontology), legal and ethical restrictions limited our access to large enough data sets. We addressed this problems in the following way.

The experiments were based on three "parent" data sets, including two data sets from the UCI repository (Asuncion & Newman, 2007) (ZOO, Cover Type) and one real-world data set (Jelovsek et al., 1989) (TOX).

From these "parent" data sets, multiple data sets were simulated by generating partial data sets.

Definition: Given a data set $D = (A, T)$, a data set $D' = (A', T')$ is called a partial data set of D if it was induced from D by selecting several columns: $A' \subset A$ and $T' \subset T$.

In total 57 partial data sets were created, including 17 for ZOO, 20 for TOX and 20 for the cover types dataset. We use partial data sets for the evaluation task to simulate a range of data sets in a limited research domain. This situation can be found again in the ACGT Project: patient data from different clinics usually intersect in their attributes.

5.2. Ontologies

For each data set presented above an ontology describing its attributes was created by hand by extracting attribute names from the data set, defining hierarchical

data set	Number of classes	number of relations
ZOO	16	15
TOX	39	38
cover	73	72

Table 1. Ontologies statistics.

relations between the attributes, and defining relations joining several attributes.

For the used data sets no other relations but hierarchical ones could be extracted. Therefore the ontology was reduced to a taxonomy. Nevertheless, the proposed characteristics can be applied to the data sets containing other relations as well. Table 1 shows some statistics of the developed ontologies.

5.3. Knowledge database

To create a knowledge database a total of five classification algorithms have been applied to 57 partial data sets. These algorithms are all based on different principles: a model-free Nearest-neighbor, a probabilistic Naive Bayes, a vector based Support Vector Machine, rules based Decision Trees and Rules Learner. The accuracy was estimated using 10-fold cross-validation.

5.4. Results

Three independent experiments, each over a different set of data sets were performed. To understand the influence of different characteristics they were divided into three groups: meta data based (D_M), graph based (D_G) and combined (D_C). For each measures group the average correlation σ^* , denoted as σ^*_{SVM} (for the SVM learning algorithm) and σ^*_{Opt} (for the distance learning algorithm), and the standard deviation (S), were estimated separately. Higher values correspond to the more correct prediction of σ_{True} and as a result can provide correct selection of the most suitable KDD process. For the performance comparison we use a default learner, estimated by a so called "take random dataset" approach, where the user chooses a similar data set randomly (labelled σ_r). Upper bounds were estimated using the training dataset with best true correlation σ_{True} (labelled σ_{max}). All values have been estimated using leave-one-out method. Table 2 sums up the evaluation results.

From these results we observe that the distance metrics learning algorithm (σ_{Opt}) produces a better prediction than SVM Learner (σ_{SVM}).

It can be also seen that proposed characteristics can effectively approximate the σ_{true}. σ_{Opt} values for ZOO and TOX datasets for the group of combined characteristic D_C are higher than for groups D_G and D_M, while for the cover data set the group of proposed characteristics D_G shows the best result. The average correlation between σ_{true} and σ^* is far more than the random default and close to the theoretical optimum. This shows that graph based characteristics indeed improve the performance.

6. Summary and Conclusions

This paper presents a new approach for dataset description in order to assist the user in selecting the most suitable KDD-process for the problem at hand. Modeling of a dataset by a graph allowed us to apply graph mining methods to perform a comparison of data sets. In total 5 topological and 6 structural properties of the graph were proposed to describe the data set. Two learning algorithms, SVM and a distance metric learning algorithm, were applied to find the correlation between developed characteristics and ranking of KDD processes.

Three independent experiments, each over a different set of datasets, show that the new approach can effectively approximate the optimal ranking of KDD processes for the given data set. The average correlation between learned and true similarity is far above the random default and close to the theoretical optimum.

Acknowledgments: The financial support of the European Commission under the project ACGT (FP6/2004/IST-026996) is gratefully acknowledged.

References

Asuncion, A., & Newman, D. (2007). UCI machine learning repository.

Brazdil, P., Gama, J., & Henery, R. (1994). *Characterizing the applicability of classification algorithms using meta level learning. machine learning-ecml94.* 83-102: Spinger Verlag.

Brochhausen, M., Weiler, G., Cocos, C., Stenzhorn, H., Graf, N., Doerr, M., & Tsiknakis, M. (June 17-19, 2008.). *The acgt master ontology on cancer - a new terminology source for oncological practice.* 21st IEEE International Symposium on Computer-Based Medical Systems, Jyvaškyla, Finland.

Cortes, C., & Vapnik, V. (Sept. 1995). *Support-vector networks.* 273-297: Machine Learning, Springer Netherlands.

Deshpande, M., Kuramochi, M., & Karypis, G. (2003). *Frequent sub-structure-based approaches for classifying chemical compounds.* 273-297: Proceedings of

		SVM		Opt.		random		max	
data set	measures	σ^*_{SVM}	S	σ^*_{Opt}	S	σ_r	S	σ_{max}	S
ZOO	D_M	0.824	0.025	0.808	0.021	0.612	0.091	0.971	0.003
	D_G	0.766	0.025	0.813	0.074				
	D_K	0.824	0.021	**0.893**	0.006				
TOX	D_M	0.743	0.021	0.760	0.031	0.51	0.054	0.929	0.007
	D_G	0.706	0.051	0.802	0.015				
	D_K	0.706	0.051	**0.815**	0.015				
Cover	D_M	0.821	0.013	0.851	0.019	0.74	0.029	0.992	0.005
	D_G	0.827	0.0075	**0.928**	0.012				
	D_K	0.831	0.0075	0.851	0.021				

Table 2. Correlations.

the Third IEEE International Conference on Data Mining table of contents, IEEE Computer Society Washington, DC, USA.

Engels, R., & Theusinger, C. (1998). *Using a data metric for offering preprocessing advice in data-mining applications.* In Proceedings of the Thirteenth European Conference on Artificial Intelligence.

Jelovsek, F., Mattison, D., & Chen, J. (1989). *Prediction of risk for human developmental toxicity: How important are animal studies for hazard identification?* 624-636: Obset. Gynecol. 74.

Kalousis, A., & Hilario, M. (2001). *Feature selection for meta-learning.* In Proceedings of the 5th Pacific Asia Conference on Knowledge Discovery and Data Mining, Springer Berlin / Heidelberg.

Kalousis, A., & Hilario, M. (Januar 2001). *Feature selection for meta-learning.* 222-233: In Proceedings of the 5th Pacific Asia Conference on Knowledge Discovery and Data Mining, Springer Berlin / Heidelberg.

Köpf, C., Taylor, C., & Keller, J. (2000). *Meta-analysis: From data characterisation for meta-learning to meta-regression.* PKDD 2000 Workshop on Data Mining, Decision Support, Meta-learning and ILP.

METAL (2002). Metal: A meta-learning assistant for providing user support in machine learning and data mining. http://www.metal-kdd.org/.

Nakano, Y., Nakamura, M., & Okabe, Y. (2007). Analysis for topological properties of the network feeding usenet news. *SAINT '07: Proceedings of the 2007 International Symposium on Applications and the Internet* (p. 14). Washington, DC, USA: IEEE Computer Society.

Neave, N., & Worthington, P. (1992). *Distribution-free tests.* London: Routledge.

Peng, Y., Flach, P.a nd Soares, C., & Brazdil, P. (2002). *Improved dataset charakterisation for meta-learning.* in Proceeding of the 5th International Conference on Discovery Science.

Sohn, S. (1999). *Meta analysis of classification algorithms for pattern recognition.* 21, 1137-1144: IEEE Trans. on Pattern Analysis and Machine Intelligence.

Todorvoski, L., & Dzeroski, S. (1999). *Experiments in meta-level learning with ilp.* 98-106: Proceedings of the 3th European Conference on Principles on Data Mining and Knowledge Discovery, Springer.

Tsiknakis, M., Brochhausen, M., Nabrzyski, J., Pucaski, L., Potamias, G., Desmedt, C., & Kafetzopoulos, D. (2008). *A semantic grid infrastructure enabling integrated access and analysis of multilevel biomedical data in support of post-genomic clinical trials on cancer.* Vol. 12, No. 2, 205-217: IEEE Transactions on Information Technology in Biomedicine, (Special issue on Bio-Grids).

Wegener, D., Sengstag, T., Sfakianakis, S., Rüping, & S., Assi, A. (2007). *Gridr: An r-based grid-enabled tool for data analysis in acgt clinicogenomics trials.* In Proceedings of the Thirth International Conference on on e-Science and Grid Computing (eScience 2007), Bangalore, India.

Xing, E., Ng, A.and Jordan, M., & Russell, S. (2003). *Distance metric learning, with application to clustering with side-information.* vol. 15, 2003.: Advances in NIPS.